CLOSING THE BUILDING ENERGY PERFORMANCE GAP BY IMPROVING OUR PREDICTIONS

A Dissertation
Presented to
The Academic Faculty

by

Yuming Sun

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Architecture

Georgia Institute of Technology
August 2014

CLOSING THE BUILDING ENERGY PERFORMANCE GAP BY IMPROVING OUR PREDICTIONS

Approved by:

Prof. Godfried Augenbroe, Advisor
School of Architecture
Georgia Institute of Technology

Dr. Pieter De Wilde
School of Architecture
University of Plymouth, UK

Dr. C. F. Jeff Wu
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Roderick Jackson
Oak Ridge National Laboratory

Dr. Jason Brown
School of Architecture
Georgia Institute of Technology

Date Approved: June 17, 2014

To my wife and our parents

ACKNOWLEDGEMENTS

I would never be able to accomplish the work without the unconditional support of my parents, my parents-in-law, and my wonderful wife Wei. Because of their love I have the belief in myself and never feel alone.

I would like to express my deepest gratitude to my advisor, Godfried Augenbroe. It has been a great honor and a joy for me to be his Ph.D. student. His guidance, encouragement, and patience help me find my passion and build up my confidence for research. I would also like to thank Professor C.F. Jeff Wu for guiding my research for the past four years. I am also grateful to Pieter De Wilde, Chris Paredis, Jason Brown, Ruchi Choudhary, and Roderick Jackson for their support of my work and enlightening conversations.

I would also like to thank all my friends at Georgia Tech who have been supportive and have made my life colorful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Increasing studies imply that predicted energy performance of buildings significantly deviates from actual measured energy use. This so-called "*performance gap*" may undermine one's confidence in energy-efficient buildings, and thereby the role of building energy efficiency in the national carbon reduction plan. Closing the performance gap becomes a daunting challenge for the involved professions, stimulating them to reflect on how to investigate and better understand the size, origins, and extent of the gap.

The energy performance gap underlines the lack of prediction capability of current building energy models. Specifically, existing predictions are predominantly deterministic, providing point estimation over the future quantity or event of interest. It, thus, largely ignores the error and noise inherent in an uncertain future of building energy consumption. To overcome this, the thesis turns to a thriving area in engineering statistics that focuses on computation-based uncertainty quantification. The work provides theories and models that enable probabilistic prediction over future energy consumption, forming the basis of risk assessment in decision-making.

Uncertainties that affect the wide variety of interacting systems in buildings are organized into five scales (meteorology – urban – building – systems – occupants). At each level both model form and input parameter uncertainty are characterized with probability, involving statistical modeling and parameter distributional analysis. The quantification of uncertainty at different system scales is accomplished using the network of collaborators established through an NSF-funded research project. The bottom-up uncertainty quantification approach, which deals with meta uncertainty, is fundamental for generic application of uncertainty analysis across different types of buildings, under different urban climate conditions, and in different usage scenarios. Probabilistic predictions are evaluated by two criteria: coverage and sharpness. The goal of

probabilistic prediction is to maximize the sharpness of the predictive distributions subject to the coverage of the realized values.

The method is evaluated on a set of buildings on the Georgia Tech campus. The energy consumption of each building is monitored in most cases by a collection of hourly sub-metered consumption data. This research shows that a good match of probabilistic predictions and the real building energy consumption in operation is achievable. Results from the six case buildings show that using the best point estimations of the probabilistic predictions reduces the mean absolute error (MAE) from 44% to 15% and the root mean squared error (RMSE) from 49% to 18% in total annual cooling energy consumption. As for monthly cooling energy consumption, the MAE decreases from 44% to 21% and the RMSE decreases from 53% to 28%. More importantly, the entire probability distributions are statistically verified at annual level of building energy predictions. Based on uncertainty and sensitivity analysis applied to these buildings, the thesis concludes that the proposed method significantly reduces the magnitude and effectively infers the origins of the building energy performance gap.

CHAPTER 1 INTRODUCTION

1.1 Importance of closing the building energy performance gap

Energy performance improvement is the central theme of sustainable building design under the pressure of global energy and environmental issues. A large portion of the overall energy savings in the United Stated will come from improving the energy performance of the built environment, both through new design and retrofit of existing buildings. Computational models that predict building energy consumption will play an ever-increasing role in the performance-based design and retrofit. In the United States, these models are applied to compliance testing against building energy codes and the assessment of the saving potential by retrofitting existing buildings (Thornton et al. 2011). For example, building energy models project that building energy efficiency can improve by 20% to 50% based on current technologies (U.S.EPA 2008); If future new technologies are taken into account, new buildings in the commercial sector could save 86% energy use with 62% of commercial buildings reaching net zero (Griffith et al. 2007).

These predictions assume that our tools are adequate to guide new designs and plan retrofits so that these savings are achieved in the 'as operated' buildings. Recent studies show that this assumption is opportunistic and overly optimistic at best (Roth et al. 2005; Turner, Frankel, and Council 2008; Ryan and Sanquist 2012). For example, a study from the New Building Institute (NBI) presented the measured and predicted energy performance for 121 LEED new construction commercial buildings (NBI 2008). Figure 1.1 shows the results. The left figure shows the predicted Energy Use Intensity (EUI) in kBtu/sf at design stage using dynamic building energy simulations against the ratio of the measured to the predicted EUI, and the right one shows the ranges of the ratios for three tiers of LEED certification. Clearly the variation of the ratios is large, and

increases with the certification tiers. Looked at differently, it can be concluded that projects with more aggressive energy performance goals are more prone to underperformance in actual energy use. It is also worthy of mentioning a modeling detail in the NBI study. As the most critical input parameters, the plug loads (which is also referred to as unregulated loads) used in predicting the results shown in Figure 1.1 were not the original values assumed by the design team. Specifically, plug loads in each model were adjusted to make up 25% of the total energy consumption in the corresponding baseline model, which was in most cases much higher than the original assumptions. The reasons for the adjustment were given in the report (NBI 2008a). Certainly, if the plug loads were modeled as they were in the original design models, the variation of the predicted EUI would be even larger, and the mean of the predicted EUI would be lower than what is shown in Figure 1.1. In summary, since the LEED system requires dynamic simulation to predict building energy consumption following the modeling guidelines of ASHRAE 90.1 Appendix G (ASHRAE 2007b), this indicates that the prediction accuracy of this modeling procedure (i.e., ASHRAE Appendix G) is low. Therefore, whether the simulation results can adequately evaluate alternative energy efficiency strategies in the support of design decision is also in doubt.

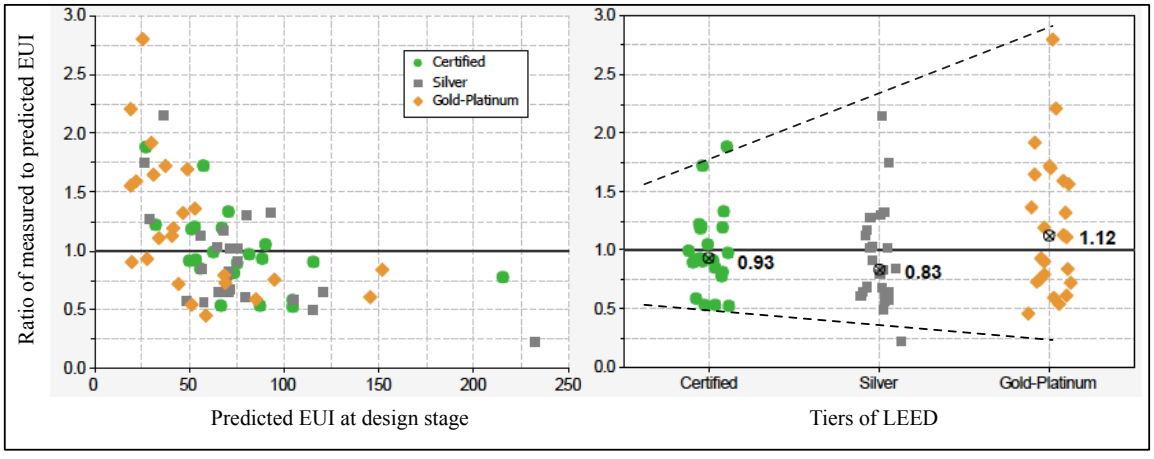


Figure 1.1 Predicted and measured EUI of LEED buildings (NBI 2008)

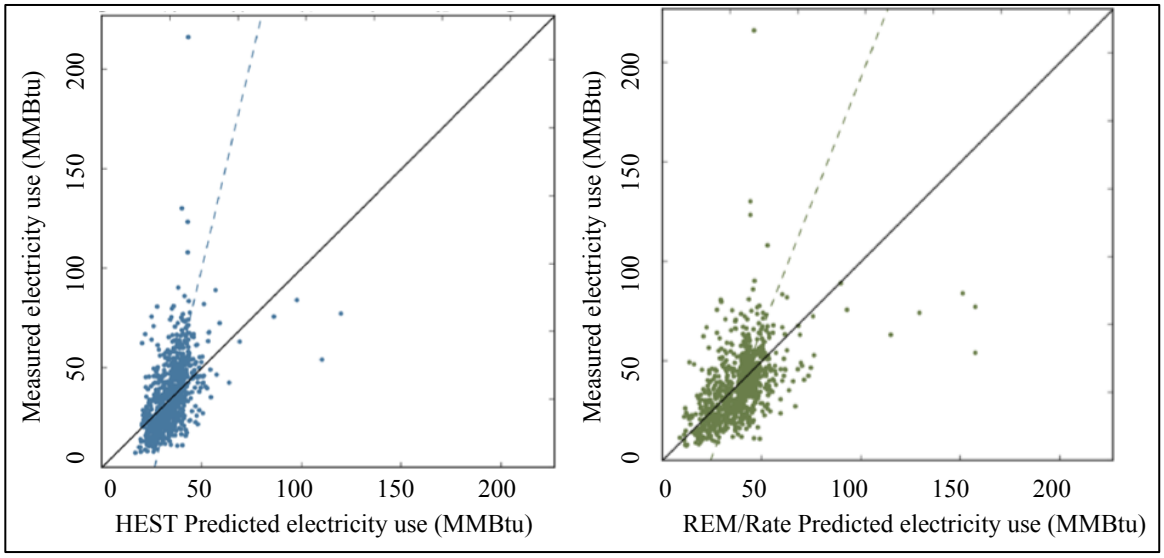


Figure 1.2 Predicted and measured electricity use of homes (Roberts et al. 2012)

In the residential domain, the situation is hardly better than what is shown for the commercial domain. In 2012, the National Renewable Energy Laboratory (NREL) conducted a study to evaluate the accuracy of different modeling systems in predicting energy performance of residential houses. Figure 1.2 shows the comparison between predicted electricity consumption with the utility billing data. The left figure evaluates the predictions from the Home Energy Scoring Tool (HEST) that is primarily developed at the Lawrence Berkeley National Laboratory, which is used to provide critical information to the Building Technology program in the Department of Energy (DoE). The right figure evaluates the predictions from the REM/Rate energy rating software developed by Architectural Energy Corporation, which is widely used in Home Energy Rating Systems (HERS). In total, 859 homes from many states in the US were included in their analysis. They showed that predictions from both HEST and REM/Rate significantly deviated from the measured electricity use. In addition, the variation in the HEST predicted electricity use over the 859 homes was much lower than the variation in the measured electricity use, whereas the REM/Rate seemed to be more consistent with

measurements in this regard. Overall, both modeling systems need research led efforts to enhance their prediction accuracy.

From the examples shown in Figure 1.1 and Figure 1.2, it is obvious that the actual energy consumption from buildings in-use can be quite different compared to what building energy models computed before the buildings are built. There are more examples by other researchers from different countries (Sunikka-Blank and Galvin 2012; Bordass, Cohen, and Field 2004; de Wilde 2014; Demanuele, Tweddell, and Davies 2010) that confirm this. This magnitude of the "energy performance gap" will certainly have serious adverse effect on a diverse group of building owners, occupants, model developers, modelers, and policy makers that is involved either by providing the predictions or by using the predictions. The 'performance gap' has also been called the 'credibility gap' by Bordass, Cohen, and Field (2004), in an attempt to raise public attention to this issue. Since 2008 it has become a legislative requirement to the publicly display an energy certificate (DEC) based on actual energy consumption in the UK (DEC 2008). Not surprisingly, the consistence between predicted building energy performances by the design team and the DEC is low (de Wilde 2014). As a result, the interest in the performance gap issue is growing within the research and industry communities (HUB 2011; Laurent et al. 2013; TM54 2013; Torcellini et al. 2006).

Even though there are many reasons for the observed building energy performance gap, including technical and administrative issues, questions about the quality of the model predictions are inevitable, given that measurements are a more direct way to observe the reality than running a computer model. Because the electricity use is relatively easy to measure with current measurement technology, the sizes of the discrepancies as shown in the figures provide no urgency to investigate unavoidable but typically much smaller measurement errors. Therefore, the wide consensus has been reached among the research and practice domain that the current model prediction capacity in understanding the actual energy consumption of either commercial or

residential buildings at their design stage is poor. Undoubtedly, people start to question the general utility of the modeling results in the support of design and technology decisions for an individual building or for multiple buildings at an aggregated level (Williamson 2010).

1.2 Why building energy consumption is difficult to predict?

Models are built to help us understand the reality with which we desire to interact for our advantage. Building energy predictions shown in Figure 1.1 and Figure 1.2 seem to poor reliability may be because the tool developers have not thought broadly enough about the applicability of the tool to real-world cases. We can further relate the discrepancies in prediction results to the lack of knowledge or data regarding: (1) reliable information about critical input parameters, e.g., lighting and plug loads; (2) modeling experience required to conceptualize building specifications with suitable model or submodels; (3) validity of the model and submodels in complicated real-world environments that could be far outside the laboratory testing conditions; (4) quality assurance of the tool, e.g. coding errors.

To better understand these many sources of errors and uncertainties, it is necessary to describe the model and its intended use along the process of design, construction, and operation. Figure 1.3 shows an example of using a computer model at the "as-built" design stage to predict "as-operated" building performance. The building energy model is used to predict the energy use intensity (EUI), or more precisely formulated, the averaged annual EUI over a number of years. At the prediction moment, a variety of model input parameters are unknown and thus are subject to significant uncertainties. For example, implementing design specifications on site by the construction team always involves a certain amount of variation in construction quality, which can also include defects or other peculiar effects in any particular case. Take sloppily installed air ducts and broken temperature sensor as examples, those types of

errors are neither regarded as a prevalent aspect of a system that is to be included in a model, nor can one predict their occurrences and effects. In addition, the operation conditions of the building are not determined at the design state. For example, the activity of the future occupants that causes plug load electricity use, and the experience of the building managers that relates to how well the building is maintained are important inputs that are impossible to predict with any certainty. Unfortunately, it has been shown that some of these factors have significant effect on the results of the quantity of interest (i.e. energy consumption). We therefore desire to predict even when we realize that the real values of these factors are impossible to be known as they still pertain to the future. Although there are some guidelines such as ASHRAE 90.1 Appendix G (ASHRAE 2007) that modelers can refer to, their published and sometimes standardized values will at best approximate the means of the unknown parameters. The approximation obviously gets better if the parameter values are categorized for different circumstances and buildings in which they may appear. However, it is still uncertain how good these standard values are for any particular building. Uncertainties in these model input parameters should contribute to a certain portion of the uncertainty in the modeling results shown in previous figures (i.e. Figures 1.1 and 1.2).

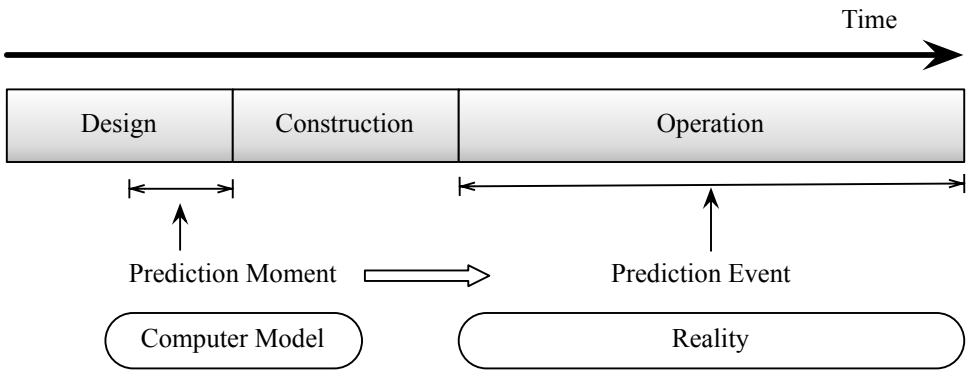


Figure 1.3 Modeling building operational performance at design stage

The second reason for the discrepancies in model predictions relates to the variability in modelers' experience. Modeling a building always involves modeler's judgment of the thermal zoning, simplification of certain architectural details, ignorance of certain physical phenomena, etc. In fact, creating building energy models from design specifications is partially a subjective art as experience has shown that even a very simple house may be described differently by different modelers (Guyon 1997). For example, looking at EnergyPlus as one of the leading modeling software in the US we find that too many modeling steps are unsupported by guidelines that predict their role in the outcome accuracy. EnergyPlus consists of a collection of modules, representing the physical process at different system scales. It is thus the modeler's decision in choosing the suitable modules for a particular building. On one hand, the flexibility can meet almost all modeling needs in a project. On the other hand, it adds considerable variability to the predictions due to the variability in modelers' experience. Where models are built to lead to conformance testing or certification levels, it is also obvious that modeling flexibility will also be prone to exploitation, potentially leading to modeler's bias in the model predictions when the modelers have other incentives than just offer their honest predictions.

In the building simulation domain, the physical-based mathematical model and code development are tightly linked. The last two sources of uncertainty (3 and 4) are strictly the concern of the model development team. However, there has not been enough research investment on model verification and validation (V&V) because most efforts are still on expending new functional capabilities. As the main V&V methods for building performance simulation, the IES BESTTEST (Judkoff et al. 2008) method is shown to be useful to detect the obvious code errors, but is much less suited to verify high-fidelity simulations against physical measurements. Specifically, depending on the cases, simulation software is regarded as equally valid if the predictions are within a reference band that is as wide as $\pm 50\%$, thus resulting in the comparable amount of uncertainty in

model predictions. Because of the importance of model validity in reducing the building energy performance gap, a more extensive discussion will be presented in the next chapter.

1.3 Goals and hypotheses

The overall objective of this thesis is twofold: (1) to build a theoretical basis for a better understanding of the building energy performance gap, and (2) to enhance our capability in predicting future building energy use, i.e. closing the building energy performance gap in a model predictive context. To achieve the goals, we not only need to adequately represent the underlying physics involved in building energy systems, but also must be able to deal with the noise and uncertainty inherent in reality. The former is indeed what most of the previous model developers aim to do, the latter, however, has only started recently.

The thesis has two major hypotheses:

Hypothesis 1: A probabilistic framework for predicting building energy performance enables us to close the energy performance gap in model predictive contexts.

Hypothesis 2: Comprehensive uncertainty quantification can offer statistically verified probabilistic predictions of building energy consumption, which will lead to substantially enhanced model predictive capabilities.

In order to procure energy efficient buildings, better information obtained from high-fidelity predictions has to be transformed into actions that lead to actual energy savings. To this end, exploratory research is carried out to inspect the role of probabilistic predictions in risk-informed design decisions of HVAC systems sizing.

1.4 Significance

The thesis adds understanding to the causes of the building energy performance gap and offers pathways towards the closure of the gap. This may also fundamentally change our thinking about predicting building energy performance in general. This research provides methods and models to characterize uncertainties in model parameters and in model forms. Uncertainty quantification (UQ) is conducted at five system scales: meteorology, urban, building, system, and occupant. With an XML based UQ repository, this work contributes to the model and data basis for uncertainty analysis at generic levels. This thesis presents methods for probabilistic prediction verification given the features and types of real-world data about building energy consumption. This research will use six buildings on Georgia Tech campus to demonstrate the value and efficacy of the proposed methods.

1.5 Thesis structure

Chapter 1 presents background and motivations for the thesis; Chapter 2 reviews the literature on the building energy performance gap and develops a theoretical basis for closing it; Chapter 3 presents methods and results of uncertainty quantification in model parameters and in model forms; Chapter 4 introduces methods for model prediction verification, especially for probabilistic prediction verification by pooling information from a collection of buildings, aiming to evaluate the validity of the underling models; Chapter 5 describes six case buildings and presents the results; Chapter 6 investigates the use of probabilistic prediction on HVAC systems sizing as an attempt to transform better predictions to informative design decisions; Chapter 7 discusses the findings, and provides conclusions and outlines for future work.

CHAPTER 2 IMPROVING THE WAY WE THINK ABOUT CLOSING THE BUILDING ENERGY PERFORMANCE GAP

2.1 Fundamental concepts

This section describes the terminology and the underlying principles of basic concepts related to predicting building energy performance. Knowing these concepts is crucial towards the formal formulation of the following research questions. What does a building energy performance gap entail? What is the role of models or empirical data in understanding the performance gap? What is the difference in formulating a building energy performance gap in a prediction context and in a post-occupancy evaluation context? Can the performance gap be avoided for new buildings at the design stage?

2.1.1 Models

Models in scientific computing are formal systems encoded from natural systems, i.e. reality (Rosen 1991). In this encoding process, there are always simplifications and approximations of reality. Specifically, the portion of the reality captured by models is an empirical-based 'enclosure' of otherwise open, interconnected natural systems (Saltelli et al. 2008). Hence, none of the models constructed as the basis for our analysis or inference are identical to the reality of our interest, or sometimes described as 'all models are wrong but some are useful' by George Box (Box and Wilson 1951). Also, building a model is as much a subjective art as it is a skill, be it that both are informed by experience. It has been shown that even a simple physical process can be described equally well by different modelers using different model formulations. This is the reason that scientists must be trained, educated, and experienced in their discipline (Burnham and Anderson 2002).

2.1.2 Model verification and validation

Verification and validation (V&V) have similar semantics and have been used interchangeably over different engineering domains due to their different history (Roache 2009; Konikow and Bredehoeft 1992). This was less an issue until numerical methods became an integral part of the outcomes of modeling and simulation. Nowadays, numerical methods are indispensable for obtaining solutions for models that entail partial different equations with complex initial and boundary conditions. Model formulation that translates reality into a set of mathematical equations requires very different knowledge than solving these mathematical equations. To distinguish these two processes, the Society of Computer Simulation (SCS) introduced formal definitions of model verification and model validation (Schlesinger et al. 1979). Figure 2.1 shows a schematic diagram describing related elements involved in V&V activities. The unidirectional inner arrows in this diagram describe the activities that translate one element into another; the outer arrows refer to the activities that evaluate the credibility of the activities. SCS introduced model verification and validation with context-specific meanings, dealing with the relationship between different entities. Verification is the process of comparing a computerized model with a conceptual model. In other words, verification no longer

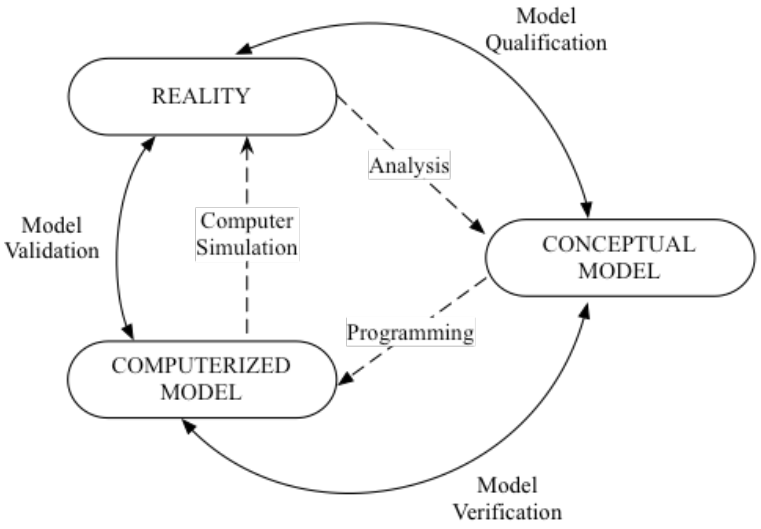


Figure 2.1 Verification and validation as they relate to reality, conceptual model, and computational model; from SCS (Schlesinger et al. 1979)

concerns any relationship with reality. On the other hand, validation is the process of comparing a computerized model (i.e., also referred to as computer model) with reality.

The definitions from SCS allow the semantic distinction between verification and validation, yet validation itself is a question of great interest to scientists at a philosophical level. In general, debates on the possibility of validation of theories have greatly enlightened scientists' view on their own profession (Konikow and Bredehoeft 1992). However, applying their debates to validation of models seems a paradox because models by definition are not the full reality, thus they are by definition not true. A slightly different perspective is that model validation is viewed as the ambition of chasing the full reality with increased knowledge regardless of whether it is attainable.

In spite of the philosophical debate on model validation, there has been a broad consensus across a variety of engineering domains from a practical perspective. In 1979, SCS (Schlesinger et al. 1979) defined validation as to substantiate that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model. In 1990, IEEE (IEEE 1990) defined model validation as a process of evaluating a system or component during or at the end of the development process to determine whether it satisfied specified requirements. In 1998, AIAA (AIAA 1998) defined model validation as a process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model. The model V&V standard in ASME (ASME 2009) and a more recent guideline from NRC (NRC 2012) adopted the AIAA's definition.

Tracing the development of model validation in engineering domains, we observe that model validation does not concern absolute accuracy any more; rather it emphasizes the utility of models in a defined application context. Therefore, these engineering communities view the objective of validation is to provide assurance or build confidence

that the models represent the reality of our interest sufficiently to guide our actions, or in particular, making decisions.

2.1.3 Prediction verification

Even though the demand on model validation is relaxed based on practical significance, it may still cause confusion regarding the meanings of assurance. In particular, models are made to provide predictions over quantities of interest that either pertain to the future or are hard to obtain from normal means. The ability of a model to reproduce what has been observed increases our explanatory abilities, but this does not mean the same quality of prediction is attainable. One important reason is that the values of model inputs are typically better known when it is used to match with historical observations than when it is used to predict a future event. To emphasize model predictive capability, the terminology of 'prediction verification' is thus introduced, such as in the weather forecasting domain (Jolliffe and Stephenson 2003). Although we think '*validation*' is more pertinent to this context than '*verification*' because it involves the comparison of predictions with measurements, we still use '*prediction verification*' to be consistent with the literature.

Another major distinction of prediction verification with model validation is that model validation typically requires well-controlled experiments. Discrepancies between modeling outcomes and measurements or benchmark data are attributable to many sources, the effects of which are naturally confounding. Model validation needs to control or eliminate all other sources of errors and uncertainties such that the remaining differences must be attributable to the mathematical model forms. In contrast, prediction verification has no intent to diagnose the causes of the differences. Its only objective is to quantify or characterize the differences. Errors and uncertainties are regarded as an integral of both the testing data and of the new model predictions, and thus are carefully processed with statistical methods.

2.2 Verification and validation in the building performance simulation domain

Building performance modeling and simulation has evolved into an indispensable area of building research over the last few decades. While efforts are still focusing on expanding new functional capabilities, less attention has been given to model verification and validation. Several organizations have commissioned large validation projects including the International Energy Agency (IEA), ASHRAE, and the European Committee for Standardization (CEN).

In the early 1980s, the IEA initiated the BESTEST (Building Energy Simulation Test) project undertaken by the National Renewable Energy Laboratory of Department of Energy (DoE) in the United States to develop a validation methodology (Judkoff, Wortman, and Burch 1983). The validation methodology is then adopted by ANSI/ASHRAE Standard 140 (ASHRAE 2004, 2007c) and the ASHRAE Handbook of Fundamentals (ASHRAE 2009b) in "Chapter 32: Energy Estimating and Modeling Methods". This validation methodology in building performance simulation was first published in 1983. Under the guidance of this methodology, a number of projects have been supported by different agencies, e.g. a four-year (2003-2007) research project, "*Annex 43 Testing and Validation of Building Energy Simulation Tools*", of the IEA Energy Conservation in Building & Community Systems (ECBCS).

Even though a lot of experiences are accumulated over these validation tests, the methodology itself remains more or less the same as in 1983. As described in the reference Judkoff et al. (2008), the ultimate goal of the validation effort is to "*investigate the ability of the simulations to predict real building performance when given accurate input data*". This validation framework consists of three tests: (1) code-to-code comparisons, (2) analytical verification, and (3) empirical validation. In the code-to-code comparison testing, a program is compared with itself or other programs. In the analytical verification testing, the outputs of a program are compared to known analytical solutions or those from a generally accepted numerical method. In the empirical validation testing,

the outputs of a program are compared to physical measurements from a real building, a test cell, or laboratory experiment. These three tests were regarded as independent, yet complementary ways to evaluate the accuracy of a whole-building energy simulation program. Another statement made in the method was that if a program successfully passed these three tests, it would be then considered validated over the test cases. This program is thus qualified to be a "certified" program, so that other programs can be tested against it.

It is not hard to conclude that the validation methodology in the building domain differs from that in other advanced engineering domains reviewed in the previous section. First, the ASHRAE V&V method permits model validation through the comparison with other models, i.e. code-to-code comparison. More importantly, the judgment regarding the agreement between model predictions and physical measurements is based on visual inspection of graphic plots. In particular, there are no quantitative conclusions about model validation whereas it still views model validation as a 'yes' or 'no' answer based on engineering needs.

The third option in the ASHRAE method, i.e. empirical validation, somewhat reflects the objective of model validation that is to approach coherence between model outcomes and physical measurements. However, it is developed from a deterministic framework, thus uncertainty is not explicitly considered. Instead, ASHRAE V&V classifies model validation experiments into three types according to the degree of control of error sources versus level of instrumentation (Judkoff et al. 2008). Class A validation experiments isolate all sources of errors, which typically involves only a few components and a few physical phenomena. Class B validation experiments control most sources of errors, and are usually conducted in a test cell (Loutzenhiser et al. 2008). Class C validation experiments do not control any error sources such that it represents buildings in operational conditions. Although class A and B validation experiments are very powerful to set up the preliminary tests of certain model assumptions at the building

component scale, these cases are usually too simple to justify the model assumptions in real buildings under real operational conditions. This research argues that understanding the efficacy of building simulation at a whole building scale requires not only model validation of components but also validation of how these components act together in real buildings.

Model validation using operational buildings needs to carefully address different sources of uncertainties and noise inherent in reality. In other words, it requires an explicit assessment in terms of the degree to which the physical experimental conditions are known. Based on this condition, it quantifies model inadequacy as the discrepancies between model predictions and physical measurements that cannot be explained by the uncertainties in the parameters and errors in the measurements. Unfortunately, an executable building energy model, dynamic simulation in particular, requires the specifications of hundreds or thousands of input parameters many of which involve considerable uncertainties under normal budget constraints that limit the study of each parameter or component in great detail. Consequently, in most cases there will not be enough confidence to attribute the overall discrepancies to the model form of the whole building energy model given the overwhelming number of parameter uncertainties.

The ultimate goal of building performance simulations is to provide reliable predictions in the support of decision-making (mostly) at the design stage or even earlier. In current practice a building simulation is routinely performed with best guesses of input parameters. However, evidence reveals discrepancies between the point predictions and the subsequent measured values (Stein and Meier 2000; Roberts et al. 2012; Turner, Frankel, and Council 2008). Although we still need more data to better understand the magnitudes of the discrepancies and the resulting effects on decision-making in different application contexts, it is for sure that building simulations are not yet capable of providing deterministic predictions (i.e. point predictions) that agree well with actual measurements.

In the building research community, we have witnessed an increasing interest in uncertainty analyses (UA) of building performance. Instead of deterministic predictions, UA provides probabilistic predictions. Such analysis is aligned with rational decision making theory and thus should lead to sound decisions; for instance risk-conscious decision making in building design and retrofit when decisions are driven by return on investment expectations, or when energy savings guarantees are part of a performance contract. Obviously, either point or probabilistic predictions should be verified based on the subsequent actual measurements. However, such a prediction verification mechanism has not yet developed in the building performance simulation domain. As a result, there is a knowledge gap in terms of assessing competing models in different prediction contexts. It is arguable that such a prediction verification mechanism will also have significant practical value because the collection and processing of large-scale energy data for real buildings in operation are attainable in the near future. This thesis will produce such a prediction verification mechanism in Chapter 4.

2.3 Predicting building energy performance

Over the last four decades, scientists and engineers have made significant progress in building energy simulation. Technically complex buildings now can be characterized with computational models that simulate hourly or monthly energy consumption for example. The role of simulation has been firmly established in the architecture, engineering and construction industry. It is used to inform decisions at scales as large as national energy policy measures and as small as the selection of shading devices of a residential house. Computer simulations suggest that a large share of energy reductions will come from introducing building energy efficiency measures (Thornton et al. 2011). The actual energy consumption from buildings in-use, however, can be quite sobering compared to what building energy models computed before the buildings are built. In many cases, buildings underperform their simulation outcomes by on average

30% and in some cases up to 100% (Sunikka-Blank and Galvin 2012; Scofield 2009; Torcellini et al. 2006; Bordass, Cohen, and Field 2004). This so-called "performance gap" has questioned the accuracy of model predictions to adequately support decisions, thus has weakened the confidence in using the computer models. A recent work by CIBSE TM54 proposes a practical guideline for assessing building operational energy consumption at the design stage (CIBSE 2013). Another prominent attempt in practice to control the performance gap is the soft landings framework (BSRIA 2009), a new paradigm for the building procurement process. However, there has not been a rigorous foundation to inspect the performance gap as an issue that can be addressed in the modeling and simulation stage.

2.3.1 Definition of the energy performance gap

The energy performance gap concept originally describes the difference between deterministic energy predictions at the design stage and the actual energy consumption during operation. This definition, however, has been altered in some recent studies where the predicted energy is no longer restricted to the design stage (Menezes et al. 2012; TM54 2013). They usually demonstrate how the calculated energy use gets closer to the actual one by replacing some assumptions of model inputs made at the design stage with more realistic information about the operational conditions, e.g., occupant schedules. The results are encouraging, but hardly informative for improving the predictions for a new design case. This is because they have circumvented the real challenge of predicting the true energy use in the presence of overwhelming uncertainties both in model input parameters and in model forms at the design stage. A well-known definition of uncertainty is given by Walker et al. (2003), who defined uncertainty as "being any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system". Based on this definition of uncertainty, it is reasonable to say that if all

uncertainties were eliminated, there would be no performance gap, only different designs. Our perspective on the performance gap is restricted to that in design predictions.

Consider the case of predicting actual energy use intensity (EUI) from given design specifications. When the model is needed to predict design outcomes, many factors that influence EUI pertain to future parameters, i.e., currently unknowable parameters, such as the operational schedules of future occupants. Figure 2.2 shows how one design specification may end up with different products that are used and maintained differently. Predicting actual EUI from design documentations involves inherent uncertainty due to the undetermined future. These inherent uncertainties result in the predicted EUI to be a random variable that should be described with probability. However, current simulation practice still favors a deterministic framework. The deterministic thinking itself leads to an incomplete representation of the reality of interest. This incompleteness makes any insights into the reasons for the discrepancy between model predicted and actual consumed energy difficult or impossible. As an alternative, an uncertainty quantification framework enables a more complete representation of possible realities using probability as a mathematical tool. Investigating

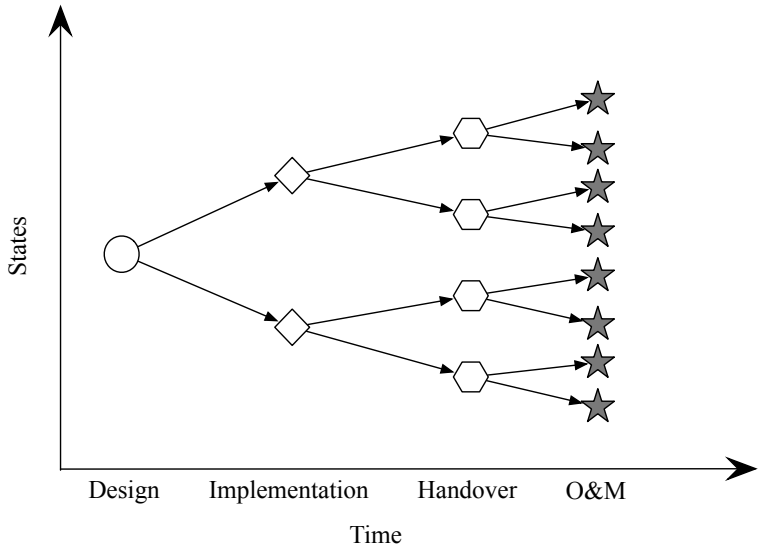


Figure 2.2 Prediction uncertainties due to unknown future

the causes for the performance gap is transformed into an uncertainty and sensitivity analysis problem.

2.3.2 Energy performance measures

Building energy performance has been communicated in different ways, at different levels of temporal and spatial resolutions. We shall specify certain quantities as measures of building energy performance to prevent any confusion due to the inconsistencies in the definitions. The definitions in the thesis are from those defined in a DoE project on the procedures for measuring and reporting commercial building energy performance (Barley et al. 2005).

With energy performance measures defined, we can relate the conceptual performance gap to measureable quantities as shown in Figure 2.3. The relations can be seen as a mapping that we use to decompose this complex problem into multiple facets. Each facet of the problem is captured by looking at a certain outcome, which we will refer to as the quantity of interest (QOI), related to the intended use of the model. In the design prediction context, the intended use of the model is characterized by certain QOI that will influence design decision-making. With the mapping as guiding principle, the

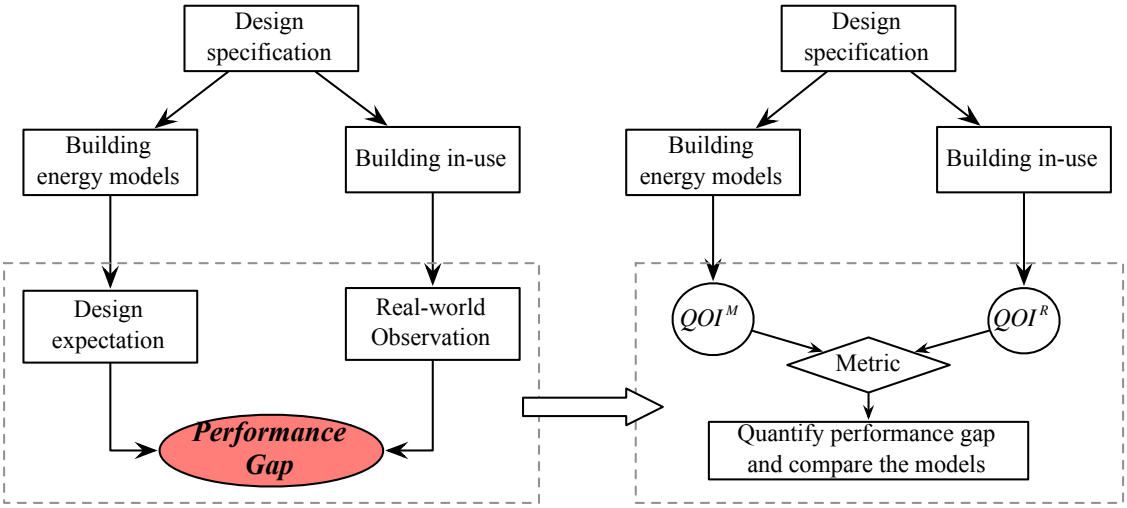


Figure 2.3 Map conceptual performance gap to measurable quantities

model is evaluated based on the comparison of model prediction, QOI^M, against real-world observation, QOI^R.

2.4 Performance gap metrics

Since the concept of '*Performance Gap*' is born under the condition of point prediction (i.e. the best estimate by the modeler), we shall introduce the basic definition coherent with this context as well. 'Gap' has the meaning of distance. Performance gap is thus defined as the absolute difference in the energy performance measure between point prediction and the real-valued physical observation. Let us take building energy use intensity as an example of a building energy performance measure to show this definition of the energy performance gap. The definition of building energy use intensity (BEUI) is given by Barley (2005) and reproduced here:

> **Building Energy Use (kWh):** Energy consumed in a building for heating, ventilating, and air conditioning (HVAC), indoor lighting, facade lighting, domestic hot water (DHW), plug loads, people movers, and other building energy use, excluding *Process Energy Use and Cogeneration Losses*. When a building has multiple functions (e.g., office, retail, laboratory, parking), the energy use may be itemized for each Functional Area for comparison to other buildings of the same types. Alternatively, the building may be analyzed as a whole and reported as a mixed-use building.
>
> **Building Energy Use Intensity** (BEUI) (kWh/m^2)
>
> BEUI (kWh/m^2) = Building Energy Use (kWh) ÷ Functional Area (m^2).

Let P be the model prediction of BEUI at the final design stage, using a Typical Meteorological Year (TMY), e.g. TMY3, or the average of predictions using a series of Actual Meteorological Years (AMY), and let R_t denote the subsequent series of realizations at different years, $t = 1,2, \dots m$. Figure 2.4 depicts the notations. The

performance gap denoted by G_t in BEUI equals the absolute difference between P and R_t i.e.

$$G_t = |P - R_t| \tag{2.1}$$

The average of G_t over a sequence of years can also be calculated by

$$\bar{G} = \frac{1}{m}\sum_{t=1}^{m} G_t = \frac{1}{m}\sum_{t=1}^{m} |P - R_t| \tag{2.2}$$

Note that Equation (2.2) has the same form as Mean Absolute Error (MAE) used to compare model predictions with observations. We shall introduce other ways of aggregating G_t in Chapter 5.

2.5 Expected performance gap in a model prediction context

Figure 2.4 shows an example of visualizing building energy performance after the model prediction period is passed and the realized values are observed. How will this be

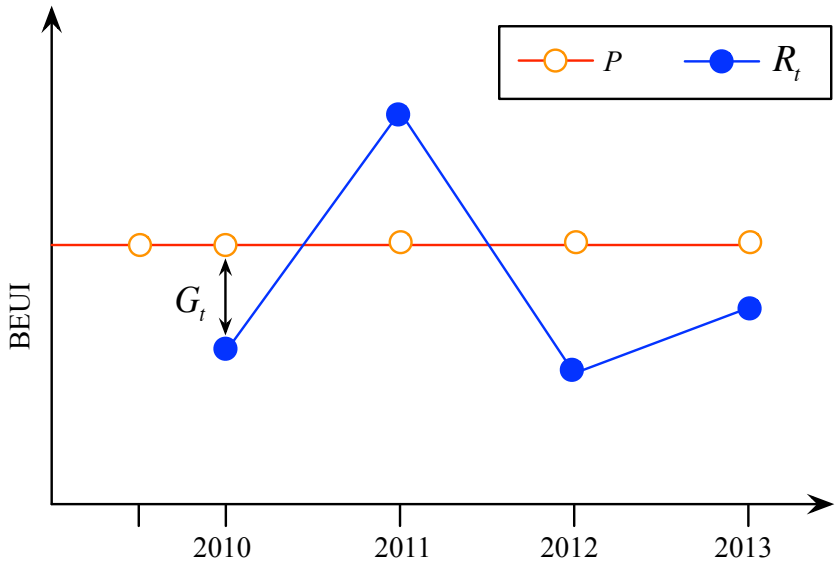


Figure 2.4 A schematic diagram of the building energy performance gap

different if we think about the performance gap in a model prediction context as shown in Figure 2.2? Let us use BEUI to explain the concept. In a model prediction context, the actual BEUI, denoted by R, pertains to a future quantity, thus should be described using probability with a probability density function $f(R)$. Figure 2.5 shows the notations of P and $f(R)$. The realized values in Figure 2.4 can be regarded as samples drawn from the distribution $f(R)$ at each year. More generally, if we allow future weather as contributing to uncertainty, the distribution of R then means the distribution of BEUI across the future years of our interest.

Since R is uncertain, the energy performance gap, G, is also uncertain. The expectation of the performance gap is computed by

$$E(G) = \int |P - R| f(R) dR. \tag{2.3}$$

Since G is uncertain, if a modeler is asked to offer a point prediction, the modeler would choose such a P that minimizes the expectation $E(G)$. For Equation (2.3), the minimum corresponds to the median of R.

To illustrate the idea, let us assume the actual BEUI is normally distributed with a

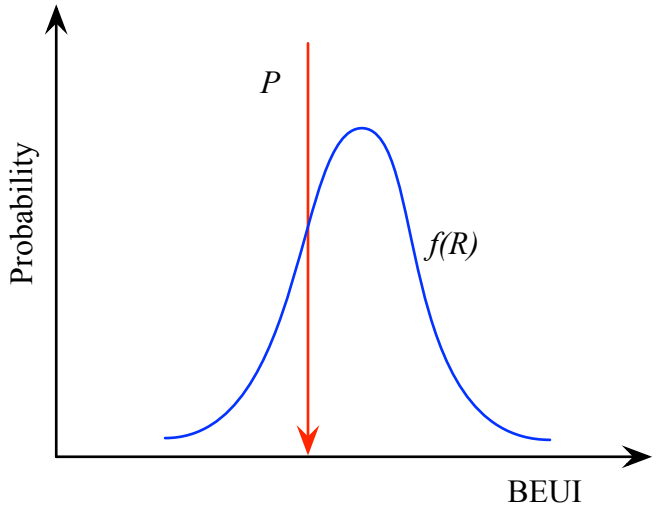


Figure 2.5 The performance gap in a model prediction context

mean of μ and a standard deviation of σ, i.e. $R \sim N(\mu, \sigma^2)$. Figure 2.6 plots the $E(G)$ against P for different distributions of R. It shows that $E(G)$ is minimized when P equals the median of R, which is also the mean of R for normal distributions. Let us suppose that the modeler is "ideal". An ideal modeler, which is also referred to as data generator, is defined as the one who can foresee the density function $f(R)$ so that a point prediction is generated that minimizes the expected performance gap. Even with ideal modeler, we can observe that $E(G)$ is always positive, the minimum of which increases with the standard deviation of R, yet is not affected by the mean of R based on how we have defined the ideal modeler. Figure 2.7 shows that the minimum of the expected performance gap increases with σ.

Note that minimizing the expectation of the performance gap does not mean that for any realized value the point prediction P always leads to the minimum discrepancy. It is possible that a non-ideal modeler wins when evaluated based on one or a few realized values by chance. However, the ideal modeler will win in the long run. This is the reason that we should not only look at the finite realized samples in our hand but also ask for the statistical confidence level to control the inference errors.

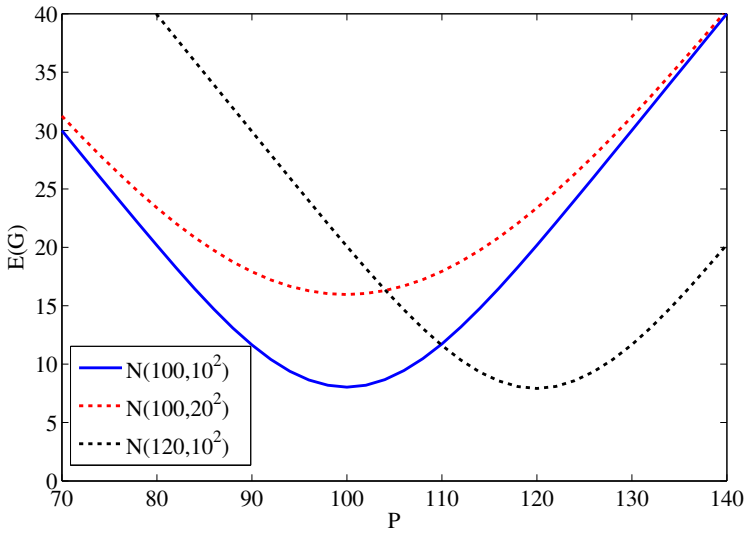


Figure 2.6 Expected performance gap, $E(G)$ versus point prediction, P

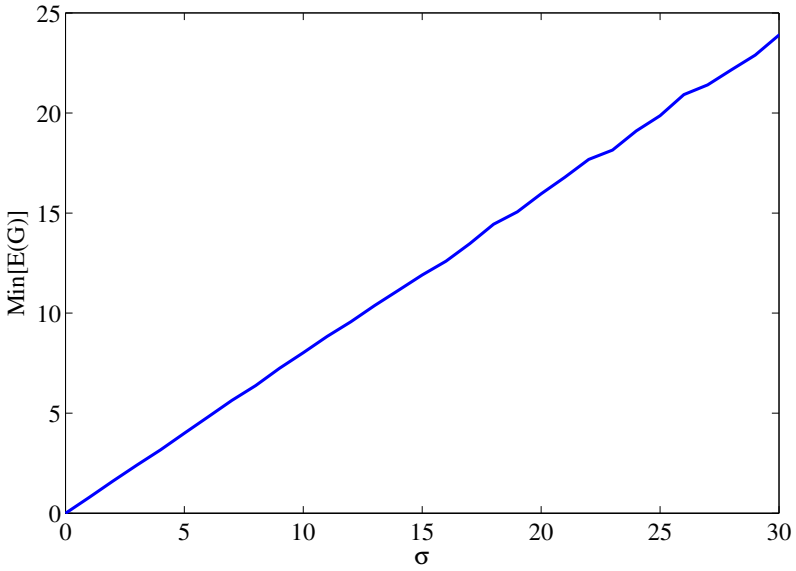


Figure 2.7 Min[E(G)] versus standard deviation σ

2.6 A formal framework for closing the energy performance gap

Now, let us first introduce a practical perspective on closing the building energy performance gap. In 2011, the zero carbon hub in the UK raised the issue in "Closing the Performance Gap: the 2020 ambition" as follows:

> "From 2020, be able to demonstrate that at least 90% of all new homes meet or perform better than the designed energy/carbon performance."

This 2020 ambition per se can be achieved by model prediction itself by offering a point prediction that is equal or greater than the 90^{th} percentile of the distribution of R. Obviously, this 2020 goal is not a good objective function because it does not encourage good model predictions to be made since it does not penalize the cases when model predictions are higher than the actual realized values. If model performance were assessed by this 2020 goal, we could simply have observed higher predictions than the actual consumption. As a consequence, there may be fewer investments in building energy efficiency because less energy savings are predicted. We certainly do not want to change the current situation where the energy efficient goal seems too high to achieve, to a new situation where we fear to set an ambitious goal in the first place.

As a suggestion, we could rephrase the 2020 ambition differently as follows: '…, be able to demonstrate that *at least 90% of all new homes are within* $\pm 10\%$ *of the designed energy/carbon performance*'. This means that the actual energy use should fall in the range of $\pm 10\%$ of the design prediction with a probability higher than 0.9. Figure 2.8 illustrates the concepts. The blue curve is the probability density function (PDF) representing the distribution of BEUI in current practice. Note, that this PDF has nothing to do with the model. It represents all possible outcomes coming from one design specification, i.e. all the states at the O&M stage shown in Figure 2.2. If the same design specification were implemented many times, the PDF would then represent the actual BEUI from these realizations. Given the PDF, we can then determine whether more than 90% of the actual realizations fall in $\pm 10\%$ of the mean. If not, we shall find ways to reduce the variation of the current PDF such that it could meet the target depicted by the future PDF. Of course, it should be kept in mind that the normal modeler has no way of knowing the PDF of all possible outcomes from an identical design specification. It is the target of this thesis to develop methods that can estimate the PDF that is statistically verified by empirical data.

In summary, we are confronting two challenges at the knowledge level on the way

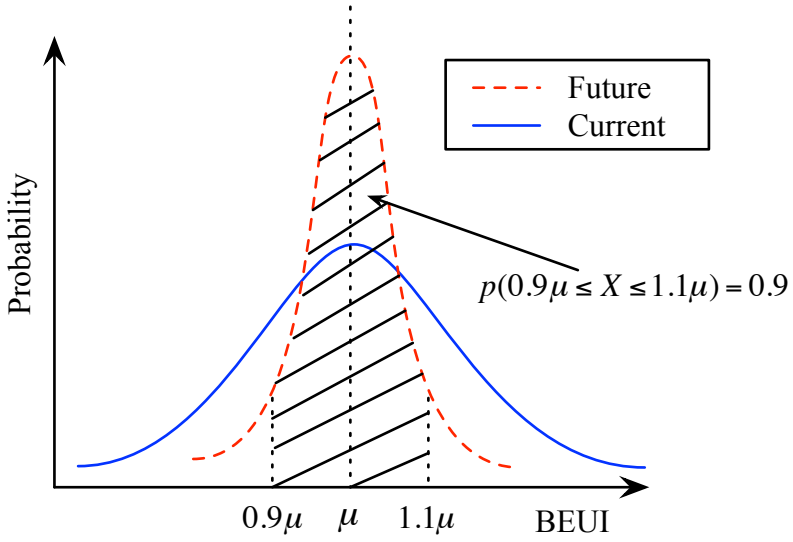


Figure 2.8 Building energy performance distribution: current status and the target

towards closing the performance gap: (1) offering the best point prediction that minimizes the expected performance gap, (2) effectively reduce the variance of actual energy use for a given design specification. It is not difficult to establish that the first part only relates to model prediction capability because it only needs the model to be a faithful representation of the current status of the industry. In other words, an optimal point prediction can always be issued as long as the modeler has the knowledge to foresee the distribution of the outcomes under the current industrial standards, i.e. all the states at the O&M stage shown in Figure 2.2. As the second part, reducing the variance of *R*, i.e. the spread of states at the O&M stage shown in Figure 2.2, cannot be attributed to the skills of the modeler. The variation of the actual energy use given a design specification would come from many factors across different stages, either controllable or not. For instance, the variability of façade leakage due to construction quality, of air imbalance due to commissioning inadequacy, and of energy waste due to operational inefficiency are factors over which we have a certain degree of control. There are also some factors on which we have little or no control, such as the weather. By saying so, there will always be a certain level of performance gap characterized by Equation (2.3) because of those inherent uncertainties. We are more interested in identifying the controllable factors whose variability leads to most of the variance in *R*. Once those factors are identified, informative advice will be offered with which the variation of the quantity of our interest will be better controlled.

Since the ideal modeler can offer the correct PDF of the outcome, he/she must also have the knowledge to correctly quantify different sources of uncertainties and their associated impacts. Although we do not have as much knowledge as the ideal modeler to completely predict the uncertainty in the future energy consumption of buildings, we at least should change our thinking regarding the representation of model prediction over a future quantity or event of interest. It has been argued in many engineering domains (Morgan and Keith 2008; Cooke 2013), for a variety of applications that the uncertainty

in the model predictions should be explicitly acknowledged, i.e. based on the probability distributions that are generated by scientific models and empirical data. For building engineering, this probability concept in model prediction not only fulfills the demand of closing the energy performance gap, but also is aligned with the natural needs of making decisions.

For example, we want to choose from two design options A and B based on their predicted energy performance. Figure 2.9 shows two ways of delivering the prediction to the designer, or the decision maker more generally. Traditionally, the BEUI is predicted as deterministic values, i.e. point prediction, as shown on the left figure. Given such a result, all decision makers will strongly prefer design option B to A in terms of their energy performance. On the other hand, if the decision maker is informed with all possible outcomes in the prediction due to uncertainties, their choice may change. The right figure shows that although design option B has a lower mean than design option A, it has considerably larger variation than design A. Because of this, there is a relative high chance that A consumes less energy than B. Therefore, the decision makers will weakly prefer B to A if and only if the expectation μ_B is lower than μ_A. Once this happens, there

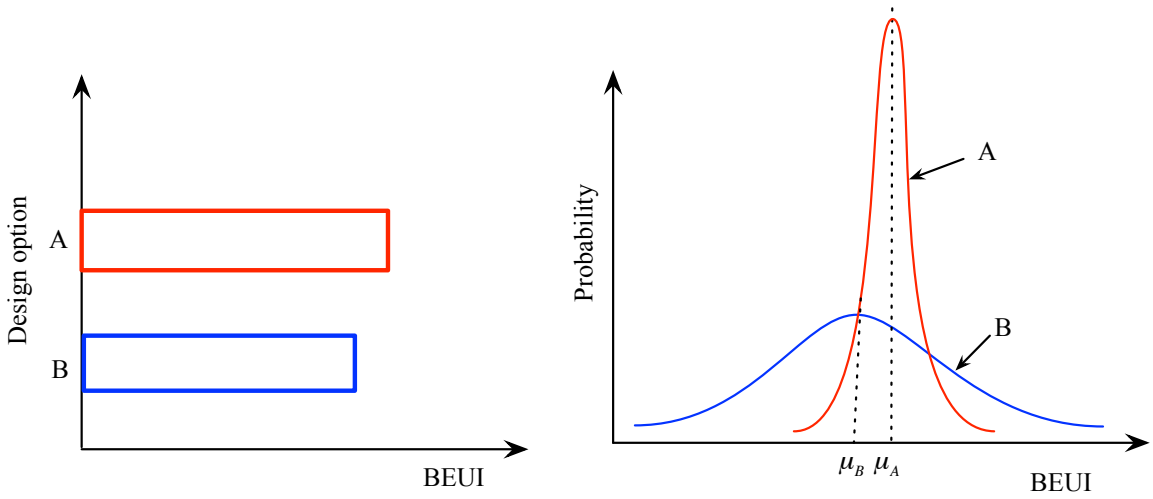


Figure 2.9 Compare A against B: deterministic prediction (left) and probabilistic prediction (right)

is no way to rank two design options such that all decision makers will agree with the ranking. Different decision makers will make the choice based on their own risk attitude once informed of the probability distributions of the outcomes.

If such situations would rarely occur in building industry then the need for probabilistic prediction would be absent and the related work would be no more than an interesting theoretical excursion. In practice though, the impact can be substantial. For example, Bordass, Leaman, and Ruyssevelt (2001) has shown the diversity of the current status of building performance in use. They categorized the complexity in buildings into four types by two factors as shown in Figure 2.10. Advanced technology has the potential to achieve more greater energy efficiency than normal technology, i.e, procure Type A buildings. However, it is also possible that the actual performance is comparable or even worse than the normal ones because of complexity in management of implementation and operation. Experience indicates that Type C buildings are not rare at all. They suggested that the building industry should be cautious in using advanced technology in that there are risks of Type C buildings associated with aggressive energy design targets accomplished by complicated technologies. Compared with our example, design B in Figure 2.9 maps to Type A or C in Figure 2.10, whereas design A maps to Type B or D.

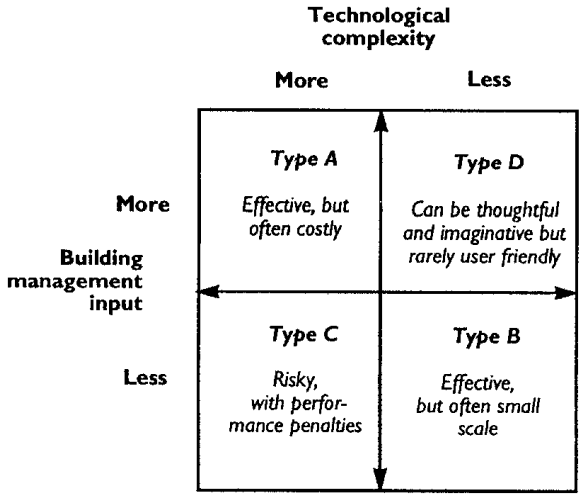


Figure 2.10 Managing complexity from Bordass, Leaman, and Ruyssevelt (2001)

Hence, our formulation can be regarded as the translation of real-world experience into a formal framework on which tangible communications among modelers, designers, professions, and building owners and occupants will be built.

However, there are fundamental issues that need extra research towards probabilistic prediction of building energy consumption, including:

- A comprehensive framework for uncertainty quantification both in model parameters and in model forms;
- A generic uncertainty quantification repository forming the basis of large-scale analysis;
- A model verification mechanism that ranks probabilistic predictions based on the consistency of predicted distributions with real-world observations generalizable to point prediction;
- An efficient probability-based sensitivity analysis that infers the major origins of prediction uncertainty.

These issues will be dealt with in the following chapters.

CHAPTER 3 PROBABILISTIC PREDICTION OF BUILDING ENERGY PERFORMANCE

3.1 Introduction

The last decade has seen a surge in the need for uncertainty analysis (UA), providing probabilistic predictions of building energy performance. The UA of whole building performance is critical in the support of risk-conscious decision-making since it informs decision-makers of underperformance risks associated with decision options. UA quantifies the magnitude of risks and their likelihoods associated with decision options. Many studies have demonstrated the significant role of UA in the context of building design and retrofit decisions. De Wit and Augenbroe (2002) obtained a probability distribution of number of hours not meeting thermal comfort to evaluate whether a mechanical cooling system is necessary. Moon and Augenbroe (2007) evaluated two remediation actions on the basis of the probability distribution of mold growth risk days. Hu (2009) evaluated the power reliability of an off-grid solar house on the basis of risk measures that reflect occupants' preferences. Recently, Heo, Augenbroe, and Choudhary (2012b) demonstrated the importance of uncertainty information for energy retrofit decision-making, especially in the context of performance-based contracts prevailing in energy service companies. These studies have shown how quantitative information about risks changes the choice of the decision option.

UA using Monte Carlo simulation considers input parameters as uncertain and propagates the uncertainties through the model by sampling from the distributions of these uncertain parameters. Hence, the UA of building performance requires two steps: uncertainty quantification (UQ) and uncertainty propagation. Figure 3.1 depicts the UA process with more details in a paper by Hu and Augenbroe (2012). It is worth noting that any model prediction has errors from many origins: physical parameter uncertainty,

model inadequacy (i.e. model form uncertainty), observation errors, and unknown longitudinal (e.g., deterioration) effects. Uncertainty in model inputs reflects the variation of parameters under partly specified and partly unknown conditions. Even if model input parameter uncertainty is ruled out, i.e., all required input parameters can be assigned the true values, the prediction will not equal the true outcome values of the process as there will always be a certain level of model inadequacy (also referred to as model form uncertainty). Observation errors account for additional discrepancies between measurements and true values. Model outcomes from a standard uncertainty analysis are presented in the form of probability density functions (PDF). In many cases an aggregation of raw model outcomes into performance indicators (PI) is done through some form of post-processing, as shown in Figure 3.1.

A few studies have quantified the uncertainty in building energy model parameters. Macdonald (2002) has quantified the uncertainty of three major types of building materials. (De Wit 2001) has summarized the uncertainty of convective heat transfer coefficients on external building surfaces. Hu (2009) has studied the uncertainty of PV module efficiency used in a zero-energy house. For existing buildings, Heo, Choudhary, and Augenbroe (2012) has applied Bayesian calibration to reduce uncertainty in the model by refining prior distributions of uncertain parameters according to

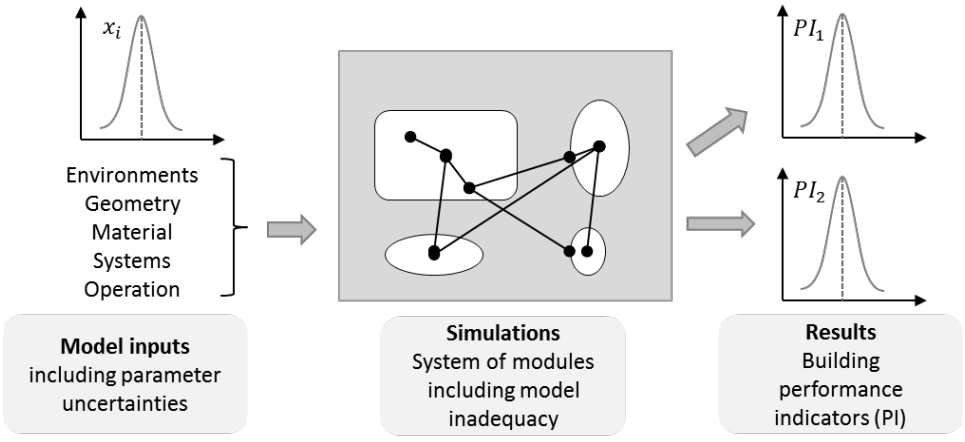


Figure 3.1 Uncertainty analysis of performance indicators (PI_i)

measured energy bills. For microclimate variables, De Wit (2001) has quantified the uncertainty in wind pressure coefficients through expert judgment for a case study building.

A comprehensive UA of BES is difficult because many factors affect building energy consumption at multiple system scales: meteorological, urban environment, building shell, building systems, and building operation. As illustrated in Figure 3.1, a BES tool is a collection of many modules with many types of interfaces among them. A number of methods of UA and SA have been developed, including differential analysis, response surface, Monte Carlo analysis, and variance decomposition. Monte Carlo (i.e., sampling-based) approaches to UA and SA are widely used because of their flexibility, ease of implementation and conceptual simplicity. Uncertainty analysis is concerned with the quantification of the uncertainty in model responses that results from the uncertainty in model inputs, whereas sensitivity analysis aims to determine the contributions of individual or a set of uncertain inputs to the uncertainty in model responses. SA ranks the importance of input parameters based on their influence on the uncertainties of model outcomes. The role of SA can be multiple. In some cases SA is used to single out the most important sources of uncertainty to start a next step in which these parameters will be better quantified. In other cases SA is used to identify the sources of risk with the aim of reducing the uncertainties by implementing better quality assurance or guarantee reduced uncertainty in performance contracts.

3.2 Literature review

The concept of uncertainty is not unfamiliar to building performance modeling and simulation. It has been acknowledged since the early 1980s when most efforts were dedicated to model development and validation. One of the best known projects was the Building Energy Simulation Test (BESTEST) (Judkoff, Wortman, and Burch 1983) in which uncertainties of model input parameters were regarded as the major sources that

complicated model validation. Although uncertainty was acknowledged in this project, BESTEST did not confront the inherent complexity of uncertainties. Instead, it emphasized the control of uncertainty up front so that uncertainty could be neglected in subsequent analysis. Uncertainties with respect to model input parameters were addressed directly in the PASSYS project (PASsive Solar Components and SYStems Testing). PASSYS adopted sensitivity analysis (SA) techniques in the model validation process (Strachan 1993) because SA can tell the relationships between model inputs and outputs that otherwise are not apparent for a complicated engineering model.

However, uncertainty quantification (UQ) of simulation results did not appear in model application contexts, such as for energy-efficient building design, before the 2000s. Some of the pioneers include Macdonald and Strachan (2001) who incorporated the UQ into ESP-r, the leading building simulation tool in Europe, and analyzed the effect of uncertainty over building design process. In 2002, De Wit and Augenbroe (2002) initiated the integration of UQ with risk analysis in a decision-making context. This study showed how a different decision would have been made for choosing between design alternatives if the decision maker were informed about uncertainties in the predictions. More recent work by Heo, Choudhary, and Augenbroe (2012) extended the application of UQ to the support of risk-conscious decision making in building design and retrofit when decisions are driven by return on investment expectations, or when energy savings guarantees are part of a performance contract.

3.3 Uncertainty characterization

The rigorous determination of the sources of uncertainty both in model parameters and in model form is a vital but often overlooked part of uncertainty analysis. To undertake this, one has to turn one's attention to a thriving area in engineering statistics that focuses on uncertainty quantification or UQ for short. This Chapter applies dedicated methods and theories that are emerging in this area of statistics to the field of

building energy models. The UQ process is organized with five systems scales: meteorological, urban, building, system, and occupants. At each level, uncertainties need to be quantified that have a potential effect on the outcomes of building energy assessment at the building and systems scales.

3.3.1 Types of uncertainty

A widely accepted definition of uncertainty was given by Walker et al. (2003), who defined uncertainty as "being any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system." Uncertainty arises in a number of ways and for a variety of reasons. It is be one of the things that are ubiquitous in virtually all fields of science and engineering.

Many uncertainty typologies are developed for many purposes (Cooke 2013; Ferson et al. 2004; Helton, Johnson, and Oberkampf 2004; Morgan 2009; Walker et al. 2003). Much of the literature divides uncertainty into two broad categories: aleatory and epistemic. The former represents uncertainties stemming from randomness, which is also referred to as irreducible uncertainty. The latter represents the uncertainty due to lack of knowledge about fundamental phenomena and is also referred to as reducible uncertainty. However, Morgan (2009), a renowned expert on uncertainty analysis in decision-making, successfully argued that the general uncertainty typology has theoretical merit but is of limited utility in applied uncertainty assessment in the support of decision-making. The primary reason is that most uncertainties involve a combination of aleatory and epistemic uncertainty. Consider the uncertainty in thermal conductivity of a wall specified in the design specifications as a simple example. Because of material and construction variability, the value of thermal conductivity is thus uncertain. The actual value can be regarded as a random sample from a population. Thus, we might categorize it as aleatory uncertainty. However, we may later realize that the thermal conductivity for this wall should be better represented as a function of temperature. The aleatory uncertainty alone

does not fully reflect our understanding of the thermal conductivity of a wall. Obviously, epistemic uncertainty is also an integral part. In building science and engineering, we believe there are few cases that can be represented as either aleatory or epistemic uncertainty. There is always a combination of both depending on the application context.

Alternatively, Morgan (2009) suggests the uncertainty in modeling and simulation is divided by (1) uncertainty about the value of empirical quantities that appear as parameters in modeling systems, and (2) uncertainty about the model functional form itself. We believe this categorization is more useful. As for building performance assessment at the design stage, model predictions differ from actual energy consumption for a variety of reasons. First of all, key model inputs such as weather conditions, building material properties, and operational schedules are usually not known with certainty or are subject to changes in real operation conditions. Models by definition ignore to some degree, and in almost all cases simplify the physical processes of the real word. Model discrepancy associated with ignorance and simplification is model form uncertainty. Other factors that account for discrepancies between model predictions and true values include measurement errors, human errors in preparing the inputs and processing the outputs. However, those are inherent obstacles (or noise) prohibiting us from observing the actual values of our interest. If possible, systematic design of experiments techniques that are well developed in statistical literature should be used to handle these factors.

The distinction between model form uncertainty and parameter uncertainty is relative to a specified modeling system, which may lead to inconsistencies across different modeling systems even if based on the same foundation for parameter and model form uncertainty. A computational model can be represented as a function $f(\bullet)$ that maps an input set $X = (x_1, x_2, ..., x_q)$ into an output $y = f(X)$. Uncertainty in X is defined as parameter uncertainty, whereas uncertainty in the function $f(\bullet)$ is defined as

model form uncertainty. The difference between these two types of uncertainty is clear yet can be blurred or miscommunicated if the function and the input parameter space have not been specified. In particular, model form uncertainty in one modeling system could be the parameter uncertainty in another modeling system or vice versa. For example, air infiltration through facades is one of many physical processes involved in modeling a full energy system of buildings. If the full building energy system accounts for the effect of air infiltration as a prescribed element of the input set X, e.g., hourly infiltration rate, the uncertainty of the infiltration rate is purely parameter uncertainty. In contrast, air infiltration itself may form a subsystem, which is explicitly modeled as a function $g(\bullet)$ with a subset of input parameters such as leakage area of exterior walls. In this case, the uncertainty of infiltration rate thus results from not only the parameter uncertainty of the leakage area but also the model form uncertainty of the function $g(\bullet)$.

3.3.2 Sources of uncertainty

Orthogonal to uncertainty typology, further categorizing uncertainty by their sources specific to a particular modeling problem is also necessary for exhaustive uncertainty quantification of complex models. Here, we organize the sources of uncertainty at five system scales as depicted in Figure 3.2. They include meteorological, urban, building, systems, and occupants. At each level uncertainties will be quantified that have a potential effect on the outcomes of the building energy assessment at the building and systems scale. Table **3**.**1** shows an example of the uncertainty matrix, providing a systematic overview of the features of uncertainty in relation to building energy predictions. The table could be used as a planning tool and be shown to the experts in the team of a UQ project. It is our experience that uncertainty occurs in almost every numerical input value and model formulation of the model and its submodels. Allocating limited resources to the most important portion of the uncertainty is essential to the success of a large UQ project. An effective method is to ask experts to

Table 3.1 Uncertainty matrix

Sources of uncertainty	Types of uncertainty	
	Parameters	**Model form**
Meteorological		Meteorological weather
Urban	Terrain roughness	Urban heat island
Building	Effective leakage area	Infiltration rate
Systems	Lighting usage density	Air distribution
Occupants	Occupant presence schedules	

preliminarily refine the candidates from a large pool of uncertainty. Table **3**.**1** provides examples of parameters or physical phenomena organized by type, scale, and sources.

3.4 Uncertainty quantification

3.4.1 Parameter uncertainty quantification

The UQ of model input parameters such as thermal conductivity of a certain type of brick is relatively straightforward. Once sufficient data are collected, the uncertainty in these input parameters can be characterized by probability distributions (e.g., normal distribution) with standard statistical methods (Bedford and Cooke 2001). Input

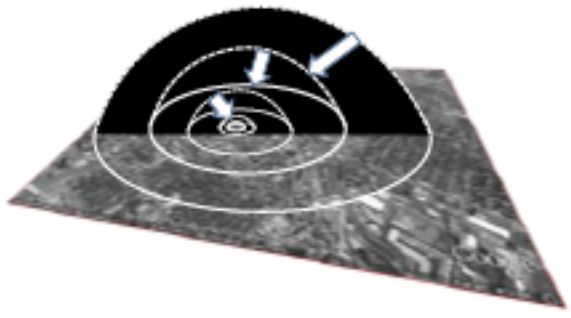


Figure 3.2 Sources of uncertainty at five system scales

parameter UQ could become difficult if the parameters are time series or if correlations are apparent among parameters. One example is the uncertainty in weather conditions. A recent paper by Lee et al. (2012) proposes to characterize uncertainty in the weather variables with a vector auto-regressive process and introduces algorithms to generate stochastic weather from historical meteorological years.

3.4.2 Model form uncertainty quantification

A model is by definition not true, i.e., a simulation result delivers unavoidably an inaccurate view of reality even if the values of the model input parameters are assigned the true values. We identify this type of uncertainty as model form uncertainty, which is also referred to as "model structural uncertainty" or "model uncertainty" in short. Model form UQ that estimates model discrepancy (also referred to as "model bias ") is a thorny problem for the modeling and simulation community (Cooke 2013; NRC 2012) despite a large methodology investment from the nuclear research sector (Oberkampf, Trucano, and Hirsch 2004; Helton, Hansen, and Sallaberry 2012; Roy and Oberkampf 2011). As for building energy simulation, model form UQ has not yet received enough attention. Therefore, we are motivated by the assumption that model form uncertainty might be an important contributor to the overall discrepancy between the predicted and the actual use of energy by a building.

Model form UQ for a complex system, such as building energy systems, is intractable at the full system scale. Moreover, it is very difficult to gather high-quality data with minimum model input parameter uncertainties at a whole system scale. An effective way of conducting UQ for complex systems is to develop a hierarchical structure that breaks down a complex system into subsystems and then into units. Figure 3.3 shows a schematic diagram of this hierarchical structure that is originally used for validation of complex modeling systems in AIAA (Oberkampf and Roy 2010). Here,

model form UQ starts from a unit problem, and then successfully moves to benchmark cases, submodel, and eventually to the complete model. Fortunately, at the unit problem level we can incorporate recent research and collect high quality physical observations. Hence, model form UQ with an ultimate goal of enhancing model fidelity should exploit this hierarchical composition, focusing first on the lowest-level component and moving successively to more complex levels. This research focuses on the submodels used for thermal load calculation for two major reasons as follows: (1) thermal load is the prerequisite of energy predictions; and (2) thermal load is more important to evaluate architectural form and fabric designs.

In current dynamic building simulation software, such hierarchy is indeed the structure of the models on which the tools are based. For example, EnergyPlus (EnergyPlus 2012) is a collection of modules that work together to calculate the final outcomes. Each module performs a specific function that involves a few physical processes. For instance, the calculation of solar irradiation on building surfaces deploys sky models formulated by a set of algebraic equations whose outcomes affect the boundary conditions of other modules. Fortunately, at the module scale we can in many

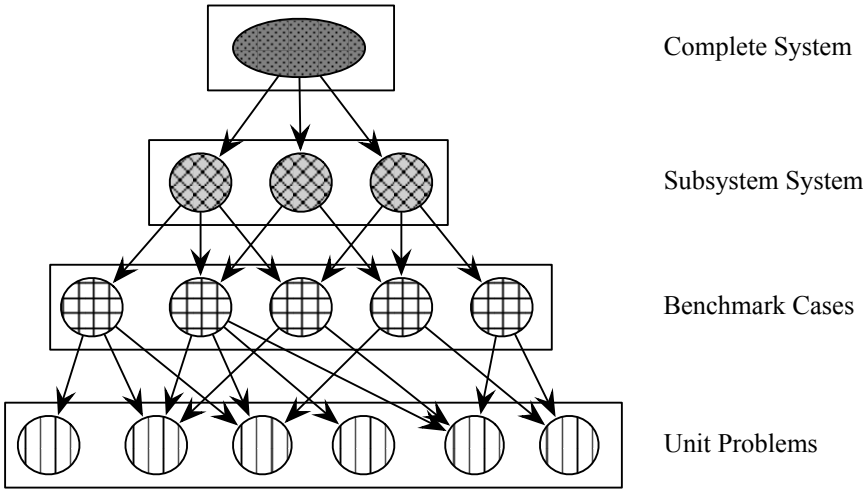


Figure 3.3 Validation hierarchy (AIAA 1998)

cases collect high quality physical observations. Hence, UQ of the building energy model should exploit this hierarchical composition, focusing first on the lowest-level components and moving successively to more complex levels.

The methods for model form UQ fall in the domain of statistical modeling and inference. Recent approaches for model form UQ are categorized into two groups, i.e., a classical frequentist approach (Oberkampf and Roy 2010; Hills, Dowding, and Swiler 2008) and Bayesian approach (Kennedy and O'Hagan 2001; Qian and Wu 2008; Chen et al. 2008). The former draws probability distributions of model bias based on statistical data analysis, whilst the latter assumes that a modeler has prior knowledge about the model bias. The experimental data are then used to update the prior distributions and obtain posterior distributions. An extensive discussion and comparison of the two approaches can be found in Hills et al. (2008). In general, when there is sufficient data available the two approaches will converge to the same result (Bedford and Cooke 2001). We use a frequentist approach for model form UQ in subsystems or units not only because sufficient data are typically attainable at these levels, but also because the frequentist approach is firmly established in the building domain and relatively easier to comprehend than the Bayesian approach.

Statistical inference for model form uncertainty quantification is based on the observations of model inadequacy under the conditions specified by the model input parameters. Because the estimate of model inadequacy falls in the category of model validation, model form UQ and validation are inherently interrelated. A discussion about their relationship can be found in a recent report (NRC 2012). In contrast to model validation that evaluates model validity for the experimental test conditions (referred to as samples), model form UQ is concerned with model predictive capability for samples that are not, or have yet to become, observable. This leads to the important foundation for the development used in this study, that model validation results are used as the

observations based on which model discrepancy and the associated uncertainty for new application samples is statistically inferred. This is explored in detail below.

Let $\eta(\boldsymbol{u})$ be the output of a computational model when the variable inputs take values $\boldsymbol{u} = (u_1, \cdots, u_k)$. Model discrepancy refers to the difference between $\eta(\boldsymbol{u})$ and the true value (but unknown) *T* under the conditions specified by $\boldsymbol{u}$ (ASME 2009). Let us consider physical observations y_{obs} that approximate the true value *T*. In this paper, we regard η as a deterministic model, i.e., $\eta(\boldsymbol{u})$ has a fixed value for a given $\boldsymbol{u}$. We can now relate model form uncertainty $diff(\boldsymbol{u}, \boldsymbol{v})$ to the simulation model outcome $\eta(\boldsymbol{u})$, and the physical observations y_{obs} in the following equation:

$$diff(\boldsymbol{u},\boldsymbol{v}) = y_{obs} - \eta(\boldsymbol{u}). \tag{3.1}$$

Note that we have added a new as yet undefined variable $\boldsymbol{v}$ to the expression of the model form uncertainty. This is motivated by the fact that the new variable $\boldsymbol{v}$ is necessary if model discrepancy displays significant correlations with some other (undetected) variables. Typically, adding new variables requires a better understanding of the shortcoming of the model at hand but relies on an effective method for physical experimental designs. As a result, adding new variables will lead to a better representation of the model inadequacy and eventually enhance model predictive capability. We take this formulation of model form uncertainty not only for computational convenience but also because the results of $diff(\boldsymbol{u},\boldsymbol{v})$ are easy to interpret since they relate to the same physical units as the model outputs. In fact, the assessment of $diff(\boldsymbol{u},\boldsymbol{v})$ is only an intermediate step. As suggested by Equation (3.1), the model output will be modified by $diff(\boldsymbol{u},\boldsymbol{v})$, so that the modified results $\eta(\boldsymbol{u}) + diff(\boldsymbol{u},\boldsymbol{v})$ will approach y_{obs}.

If the physical observations are not attainable, it is also possible to replace y_{obs} by model outcomes obtained with a high fidelity model. An application of this approach is shown in (Sun, Heo, et al. 2014) for the UQ of building microclimate variables. This method presumes the existence of a high fidelity model whose model inadequacy is of secondary order effect compared to that of the low fidelity model under study. Given the features of building simulation, we found it was useful to explore the high fidelity model as an option to quantify the model form UQ of the low fidelity model. As aforementioned, a complete building energy model consists of many submodels. A submodel that has a reduced order implementation in the building energy model could also exist in a higher order implementation, e.g. to deliver the primary quantity of interest in another domain. For example, Sun, Heo, et al. (2014) used a high-order meteorological model as the high fidelity model to quantify the uncertainty in a reduced order model of building microclimate.

3.5 Examples of parameter uncertainty quantification

This section introduces methods of parameter UQ using examples. We start from a simple case where UQ deals with a single parameter – ground albedo. The second example, convective heat transfer coefficients, deals with two parameters with correlation. Finally, we use lighting and plug load as the third example to illustrate more complicated parameter UQ.

3.5.1 Example 1: ground albedo

A monthly reflectance of 0.2 is embedded in the standard calculation as the default value and is commonly used in practice. To quantify the uncertainty in this parameter, we consider ground as composed of impervious and pervious road compositions, the aggregated ground albedo is calculated by Equation 3.2:

$$\rho_{ground} = \rho_{prvd} f_{prvd} + \rho_{imprvd}(1 - f_{prvd}), \tag{3.2}$$

where ρ_{prvd} , ρ_{imprvd} are pervious and impervious road solar reflectance, f_{prvd}, f_{imprvd} pervious and impervious road area fraction. The range of each parameter comes from the global dataset and references (Ahrens 2007; Clarke et al. 2002) and is summarized in Table 3.2.

The uncertainty in each of the parameters is modeled with independent uniform distributions. For each terrain type, the probability density function (PDF) of ρ_{ground} is estimated by generating a Monte Carlo sample and utilizing this sample to construct a Kernel density estimator. Figure 3.4 shows the estimated PDF for city terrain. The distribution is centered at around 0.25, and it has a range from about 0.05 to 0.45.

As ground reflectance increases dramatically in the presence of snow, the chosen reference energy model, EnergyPlus, incorporates a snow modifier to modify ground reflectance for snow condition. The UQ of ground reflectance in the presence of snow is based on a review of the literature. Snow reflectance can vary from 0.75 to 0.95 for fresh snow cover and 0.4 to 0.7 for old snow cover (Muneer 2004). For urban area, a research by Hunn and Calafell (1977) concluded that no characteristic ground reflectance in winter could be specified and suggested values ranging from 0.16 to 0.49. For the rural area, they suggested ground reflectance of 0.6 to 0.7 for most rural landscape where a large snow cover is visible without obstruction.

Table 3.2 Experimental design setup for ground reflectance

Parameters	**Large city centers**		**Urban and suburban areas**		**Open country**	
	Min	Max	Min	Max	Min	Max
f_{prvd}	5%	25%	5%	90%	90%	1
ρ_{imprvd}	0.072	0.44	0.072	0.44	0.072	0.44
ρ_{prvd}	0.05	0.4	0.05	0.4	0.05	0.4

3.5.2 Example 2: convective heat transfer coefficients

Convective heat transfer at internal and external building surfaces has been regarded as one of the most determining, yet hardest to determine factors in heat balance calculations, significantly influencing the prediction of building energy consumption. This is due to the non-linear and complex spatial and temporal variability of the convective heat transfer that occurs in a given configuration. The large variation in h_c correlation models derived from field measurements and CFD simulation is an indication of the complexity of the underlying physical phenomenon. It appears that no single correlation model of h_c is predominantly superior to the others. We should consider all these empirical study as equally likely in estimating the actual h_c over building surfaces. Thus, convective heat transfer coefficients for both external and internal surfaces are quantified by reviewing models and values reported in published literature and standards.

For example, the external convective heat transfer coefficient, h_c, uses wind speed as the regression variable and typical takes the following form:

$$h_c = aV_z + b, \tag{3.3}$$

where V_z is the local wind speed, and a and b are empirical coefficients. A number of field tests have been carried out to derive these two coefficients based on their own independent onsite measurements. The values of these two coefficients vary from one

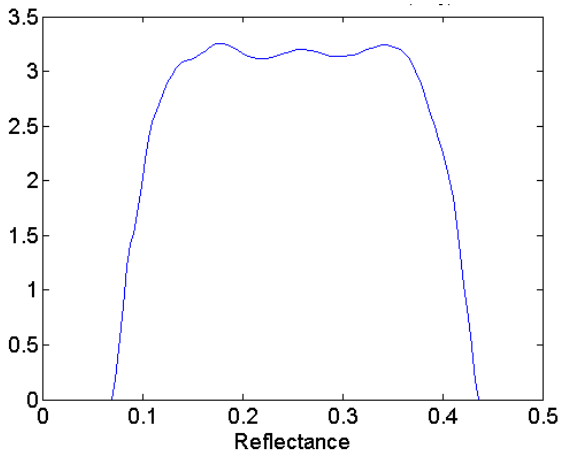


Figure 3.4 PDF of ground reflectance for large city centers

study to another. Table 3.3 summaries the h_c relationships developed the literature since 1970. To visualize the uncertainty in the calculation of h_c, Figure 3.5 plots the values of h_c calculated from each relationship against wind speed. It shows the uncertainty in the calculation of the h_c is drastic. For a given level of wind speed, the maximum value can be 10 times larger than the minimum value.

We quantify the uncertainty of h_c through the two regression coefficients a and b such that its functional form with wind speed can be retained. Because a and b always appear as a pair in the literature, we shall use bivariate normal distribution to retain their correlation structure.

We explore bivariate kernel density estimator to estimate the density of the two coefficients. The bivariate kernel density estimator takes the following form:

$$\hat{f}_H(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_H(\boldsymbol{x} - x_i) \tag{3.4}$$

Where $\boldsymbol{x}_i = (a_i, b_i)^T$, $i = 1,2,...,n$ are the indices of observation pairs, K is the kernel function, and H is the bandwidth 2×2 matrix, and K_H is given by

$$K_H(\boldsymbol{x}) = |H|^{-\frac{1}{2}} K(|H|^{-\frac{1}{2}} \boldsymbol{x}) \tag{3.5}$$

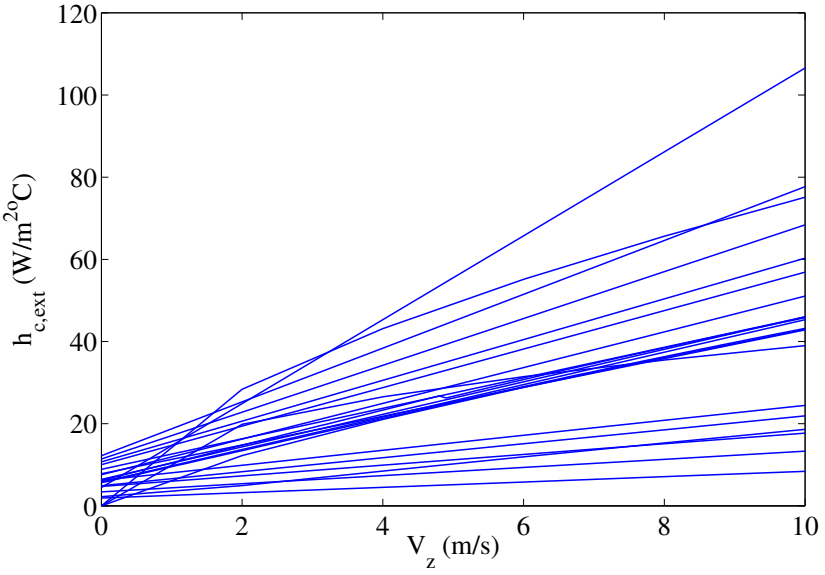


Figure 3.5 External convective heat transfer coefficients uncertainty

Table 3.3 External surfaces h_c relationships in the literature

Models	Experimental conditions	References
$h = 3.50V_w + 7.82$	Wind tunnel, very smooth surface	(Jennings 1970)
$h = 3.71V_w + 8.9$	Wind tunnel, smooth surface	(Jennings 1970)
$h = 4.96V_w + 10.7$	Wind tunnel, cast concrete and smooth brick	(Jennings 1970)
$h = 18.65V_w{}^{0.605}$	Field measurements	(ASHRAE 1975)
$h = 5.7V_w + 11.4$	Field measurements	(Cole and Sturrock 1977)
$h = 1.824V_w + 6.22$	Field measurements, window surface	(Kimura 1977)
$h = 4.35V_f + 7.55$	Field measurements,, window surface	(Nicol 1977)
$h = 1.3V_w + 4.7$	Field measurement, 18th floor center	(Sharples 1984)
$h = 1.7V_w + 4.9$	Field measurement, 18th floor edge	(Sharples 1984)
$h = 0.99V_w + 3.4$	Field measurement, 14th floor center	(Sharples 1984)
$h = 0.65V_w + 1.9$	Field measurement, 6th floor center	(Sharples 1984)
$h = 2.38V_w{}^{0.89}$	Field measurement, window, low-rise buildings	(Yazdanian and Klems 1994)
$h = 14.82V_w{}^{0.42}$	Field measurement, 6th floor vertical surface	(Taki and Loveday 1996)
$h = 4.69V_w + 10.03$	Laboratory measurements,	(Kumar et al. 1997)
$h = 6.548V_w + 12.2$	Laboratory measurements,	(Kumar et al. 1997)
$h = 3.91V_w + 5.62$	Wind tunnel, < 4.88 m/s, smooth surface	(Clarke 2001)
$h = 7.17V_w{}^{0.78}$	Wind tunnel, < 30.48 m/s, smooth surface	(Clarke 2001)
$h = 4.28V_w + 6.19$	Wind tunnel, < 4.88 m/s, rough surface	(Clarke 2001)
$h = 7.6V_w{}^{0.78}$	Wind tunnel, < 30.48 m/s, rough surface	(Clarke 2001)
$h = 10.21V_w + 4.47$	Field measurement, on vertical wall	(Hagishima and Tanimoto 2003)
$h = 3.95V_w + 5.8$	Wind tunnel, plate, parallel flow, < 5 m/s	(Palyvos 2008)
$h = 7.13V_w{}^{0.78}$	Wind tunnel, plate, parallel flow, < 5 m/s	(Palyvos 2008)

We use a standard bivariate normal density as the kernel function K. H is determined by optimization to be:

$$H = \begin{pmatrix} 0.52 & 0.50 \\ 0.50 & 3.30 \end{pmatrix} \tag{3.6}$$

3.5.3 Example 3: lighting and plug load uncertainty quantification

3.5.3.1 Lighting and plug loads in building energy models

Lighting and plug loads are not only the important electricity consumers but also are major sources of the cooling load in buildings. Both lighting and plug loads vary over the simulation time, and thus need to be treated as time series with hourly time resolution for most dynamic simulations. Given the limited information about building operation at the design stage, lighting and plug loads are rarely directly entered by the modeler as an 8760-hour time series. In most cases, the hourly values of lighting and plug loads are derived from the hourly peak use intensity and three daily profiles including weekday and weekend. The elements in these profiles are the ratios of the lighting or plug loads of each hour to the peak value. These profiles is thus also referred to as diversity factors, representing the hourly variation of the lighting or plug loads. A whole-year profile is constructed from the three representative daily profiles used repetitively. The hourly lighting and plug loads for a representative day can be expresses by the following equation:

$$y_t = q \times d_t, t = 1,2,\cdots,24 \tag{3.7}$$

where q is the hour peak use in W/m^2 over a year, d_t is a hourly profile of the representative day. The UQ of y_t is thus decomposed into the UQ of q and d_t.

3.5.3.2 Data collection

The data used to conduct the UQ comes from the ASHRAE Research Project 1093-RP (Abushakra et al. 2001; Claridge et al. 2004). This project collects field data that include hourly measurements of lighting and plug loads for office buildings. The objective of the project was to develop typical lighting and plug load hourly profiles in office buildings. In the work, we will process their dataset, aiming to quantify the

uncertainty of these profiles besides the typical ones. There are two major types of data: lighting and plug loads are separately monitored, and the summation of lighting and plug loads is monitored. In addition, the data quality from each monitored building is classified as good, ok, or unknown.

3.5.3.3 UQ of peak use density, q

We first show a boxplot including all data in Figure 3.6. L G1 denotes the lighting peak use from 8 buildings in Group 1; PL G1 denotes the plug loads from 8 buildings in Group 1, L+PL G1 denotes the total of lighting and plug loads using 7 buildings with separately monitored lighting and plug load in Group 1; L+PL G2 denotes 23 buildings with lighting and plug loads monitored as an integral. From this plot, it is quite obvious that the total lighting and plug loads from buildings in Group 1 is substantially lower that that from buildings in Group 2. Given that Group 2 contains a larger number of buildings, buildings in Groups 1 may not be able to represent the distributions of lighting and plug loads for office buildings. This becomes more evident if we remove the unknown quality of data points. There are only two data points left for lighting separately monitored (Figure 3.7), one data point with plug loads separately monitored, and 24 data points in Group 2. Thus, this dataset will not allow the lighting and plug load to be separately quantified.

Before the uncertainty quantification of the total lighting and plug loads, we explore whether there is any significant correlation with building size. As Figure 3.8 shows, no pattern seems visible between the two parameters. To be rigorous, we also compute the correlation coefficient and conduct a hypothesis testing. The result confirms that the correlation between the two parameters is not statistically significant at the level of 0.05 ($p = 0.27$).

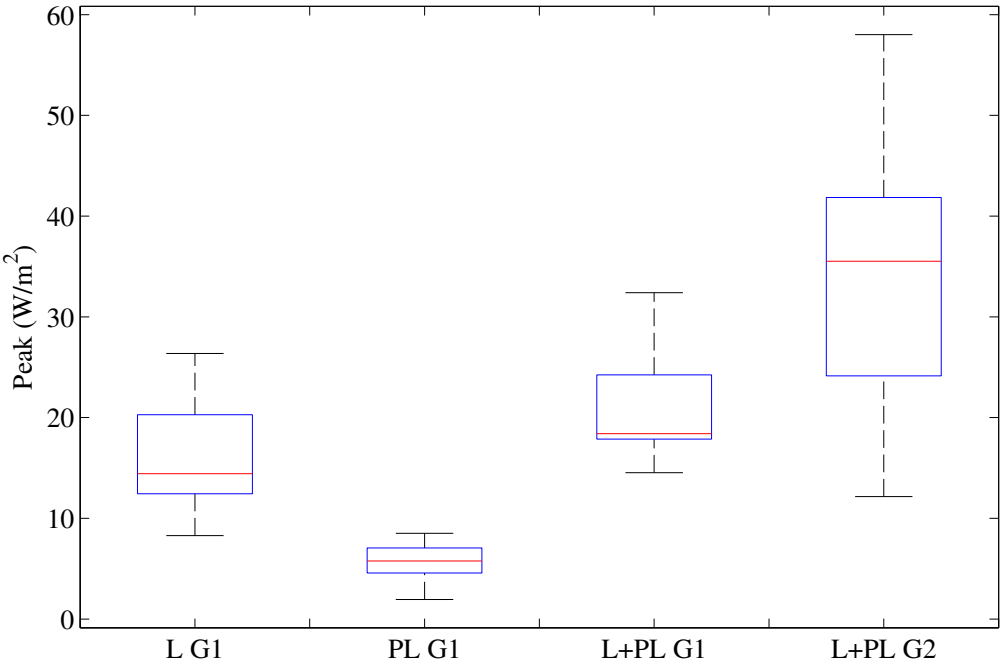


Figure 3.6 Peak use boxplots including data with all levels of quality

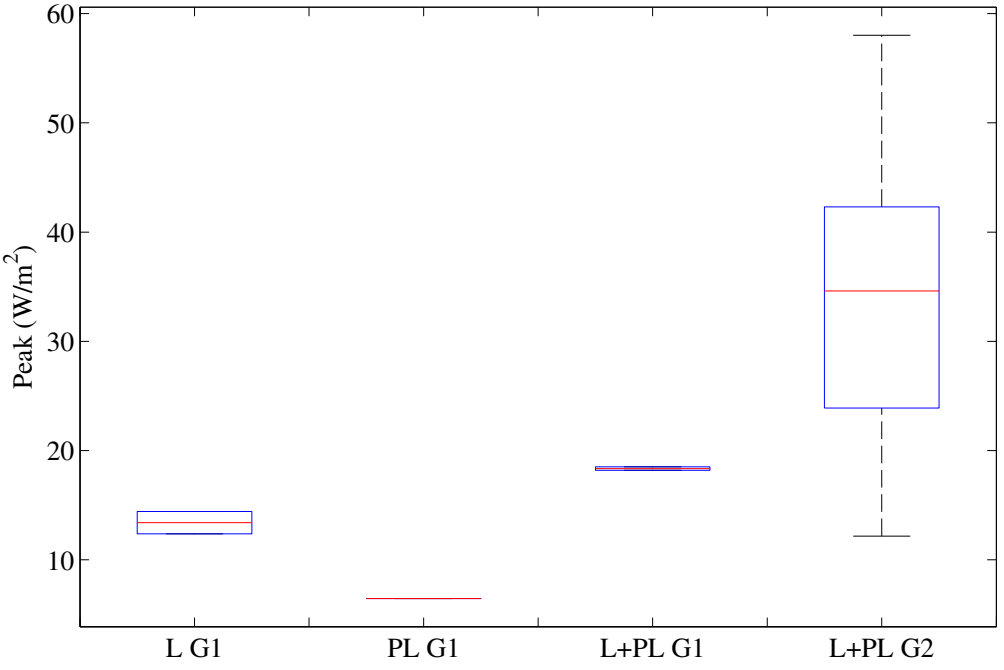


Figure 3.7 Peak use boxplots including data quality assessed as good or ok

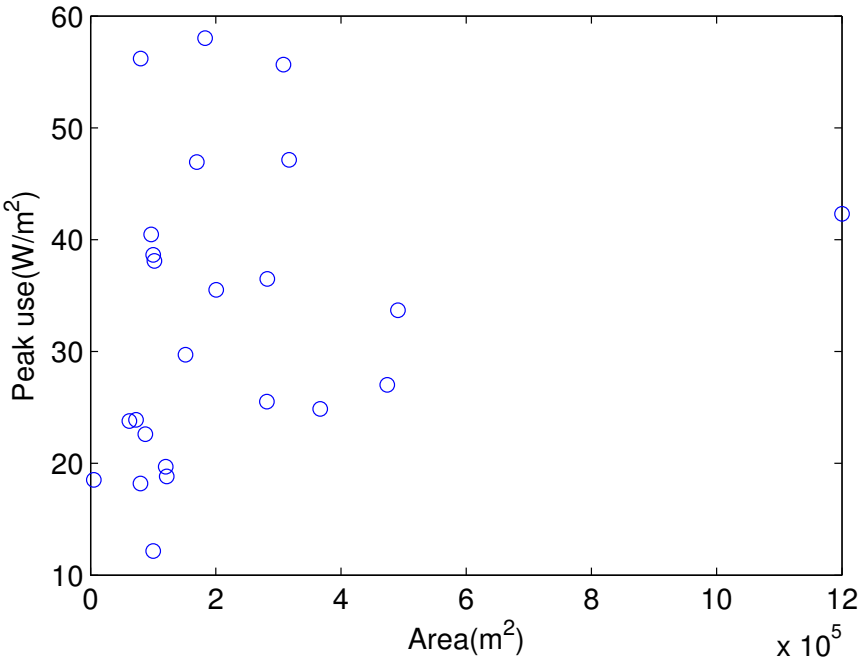


Figure 3.8 Peak use of lighting and plug loads vs. building area

Finally, the uncertainty of the total peak use of lighting and plug load is characterized as a Gaussian distribution. Figure 3.9 plots the empirical CDF of the peak use and the CDF of the normal distribution, and shows the results of the K-S test, a statistical hypothesis test whose null hypothesis is that the peak use is normally distribution. The K-S result fails to reject the null hypothesis at the significance level of 0.05 ($p = 0.774$). Therefore, we shall use the normal distribution to fit the uncertainty of this parameter. Thus, the peak use of lighting and plug load is characterized as a normal distribution with a mean of 33.1W/m^2 and a standard deviation of 13.1W/m^2, i.e. $N(33.1,13.1^2)$.

3.5.3.4 UQ of diversity factor d_t

In the ASHRAE project, the diversity factors are reported for weekday and weekend at the building level. Figure 3.10 plots the hourly diversity factors of weekday and weekend; each curve comes from one office building. Besides the hour-to-hour variation for a given building, which reflects the building operation profiles during work

hours and off-work hours, the building-to-building variation is also quite large, meaning that different office buildings operate differently. This variation is more evident during off-work hours than the working hours. For example, the diversity factor during the off-work hours (22:00 to 3:00) ranges from 0.3 to 0.7 among different office buildings. This weekend diversity factors also manifests the building-to-building variation.

Moreover, we detect large correlation between weekday and weekend diversity factor. We analyze their correlation in two aspects. We firstly calculate the correlation coefficient between averaged diversity factor for weekday, $\mu(d_t):weekday$, and weekend, $\mu(d_t):weekend$. Figure 3.11 plots $\mu(d_t):weekday$ against $\mu(d_t):weekend$. The correlation coefficient is 0.941, meaning high linear correlation between the two factors. We also analyze the correlation between the last hour of weekday and the first hour of weekend. Their correlation coefficient is also as large as 0.975. Intuitively, this means that if a building has high diversity factor during weekdays, it also has high diversity factor during weekends.

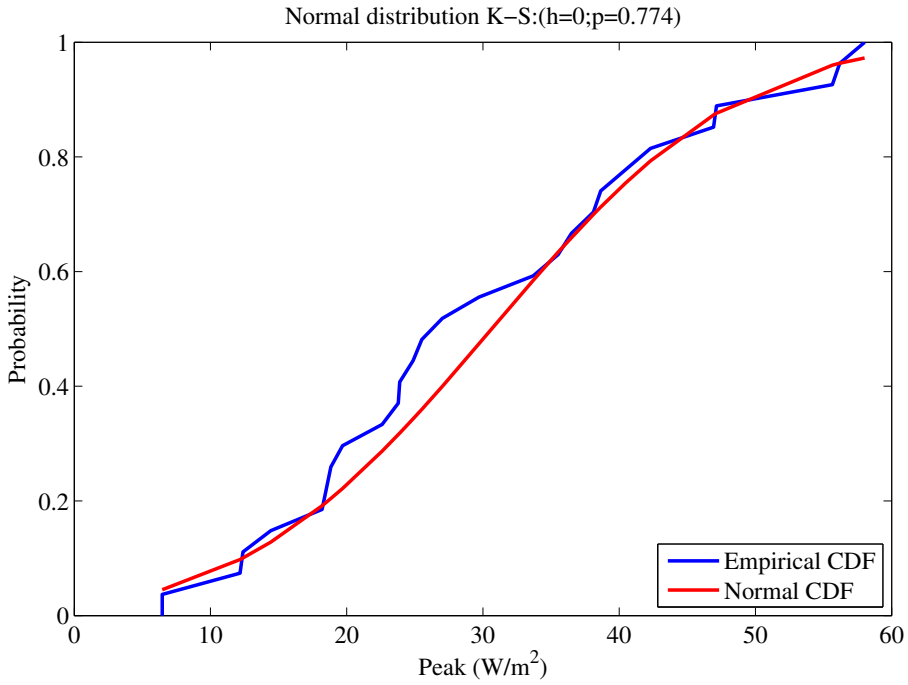


Figure 3.9 Normal distribution of the peak use of lighting and plug loads

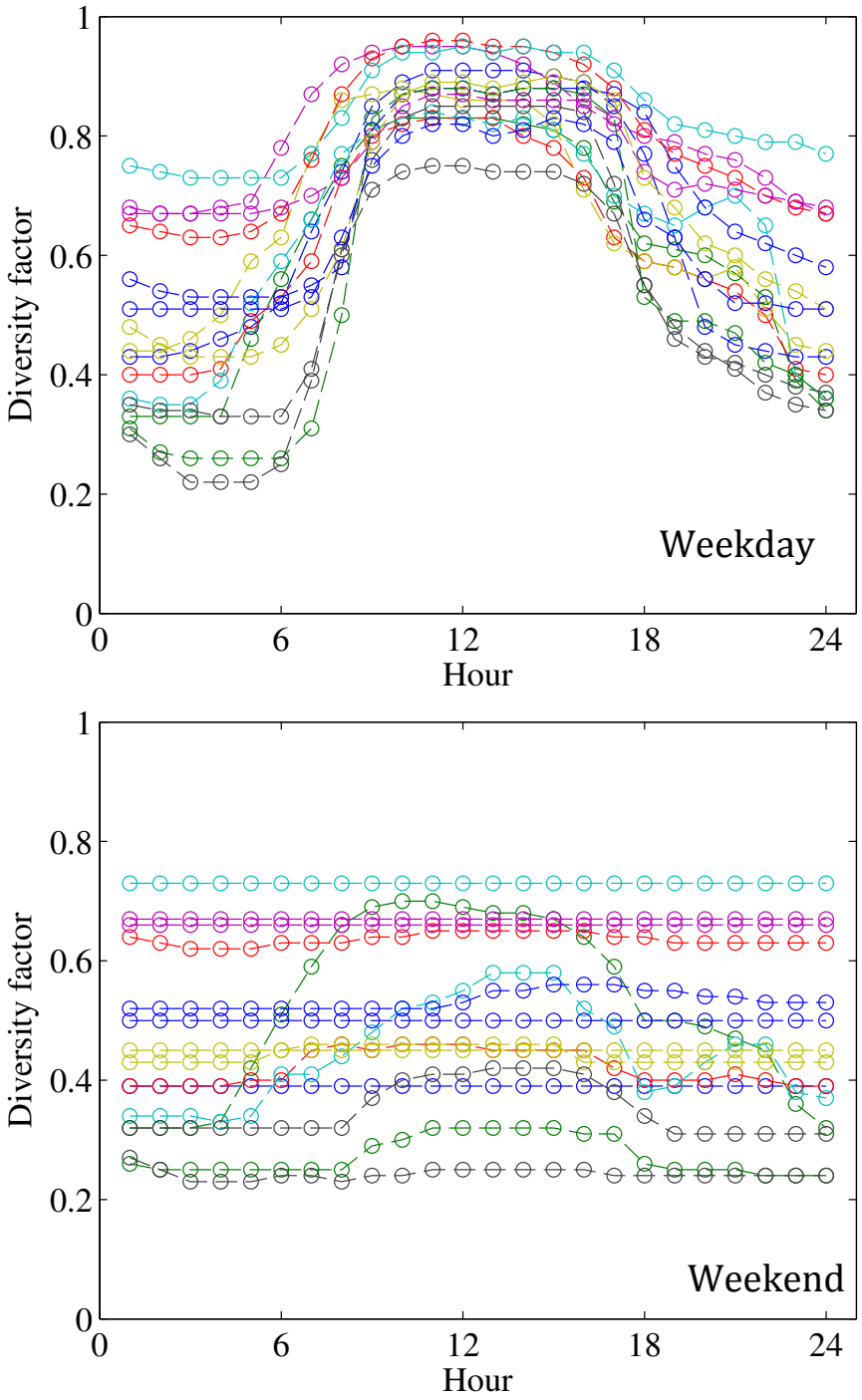


Figure 3.10 Hourly diversity factor profiles for weekday (a) and weekend (b)

Given the correlation between weekday and weekend, it is problematic to treat two sets of diversity factors independently. To consider the correlation, we combine weekday with the weekend, such that a 48-dimension vector is used to represent the

profile of the diversity factor for each building. After certain pre-processing the raw data, we obtain 15 buildings, so that we have a data matrix of 48×15. We intend to derive a distribution for a 48-dimension vector using this dataset. In this case, we use a multivariate normal distribution to fit the data. It is straightforward to calculate the means from the averages by column. In terms of the 48×48 covariance matrix, the data size at hand is not sufficient to perform regression analysis. Instead of using the full covariance matrix determined by the data, we conducted two trials, which are commonly used when data is lacking, in terms of the covariance matrix.

The first trial is to assume the correlation between the any two adjacent hours is one, whereas the correlation between any two non-adjacent hours is zero. Figure 3.13 shows 20 random samples generated with the assumption. The second trial is to use a tridiagonal covariance matrix. The correlation between any two adjacent hours is given by the data whereas the correlation between any two non-adjacent hours is zero. Figure 3.14 shows 20 random samples generated with the second trial. The first trial generates all parallel diversity factor profiles that are bounded by the range of the observed 15 profiles of the diversity factor. The second trial using the tridiagonal covariance matrix can generate some stochastic patterns from hour to hour that seems more plausible than the first trial. However, among the 20 samples from the second trial, we observe a few samples in the weekend whose nighttime diversity factors are higher than those during the daytime, which is unlikely in reality.

Moreover, the objective here is to generate a range of plausible weekend and weekday diversity profiles that can capture the variation due to uncertain operations (building-to-building variation), and meanwhile obey the general pattern of building operation during work hours and off-work hours. Given the limited data, it is one's judgment in preferring one to the other. Given that the primary sources of uncertainty is the building-to-building variation in terms of the operational conditions. It is expected that two different covariance matrices will result in similar impact on the uncertainty of

evaluated as monthly or annually aggregated building energy consumption. Here, we choose the first trail for its simplicity. When more data are collected in the future, we will refine the UQ of the diversity factor.

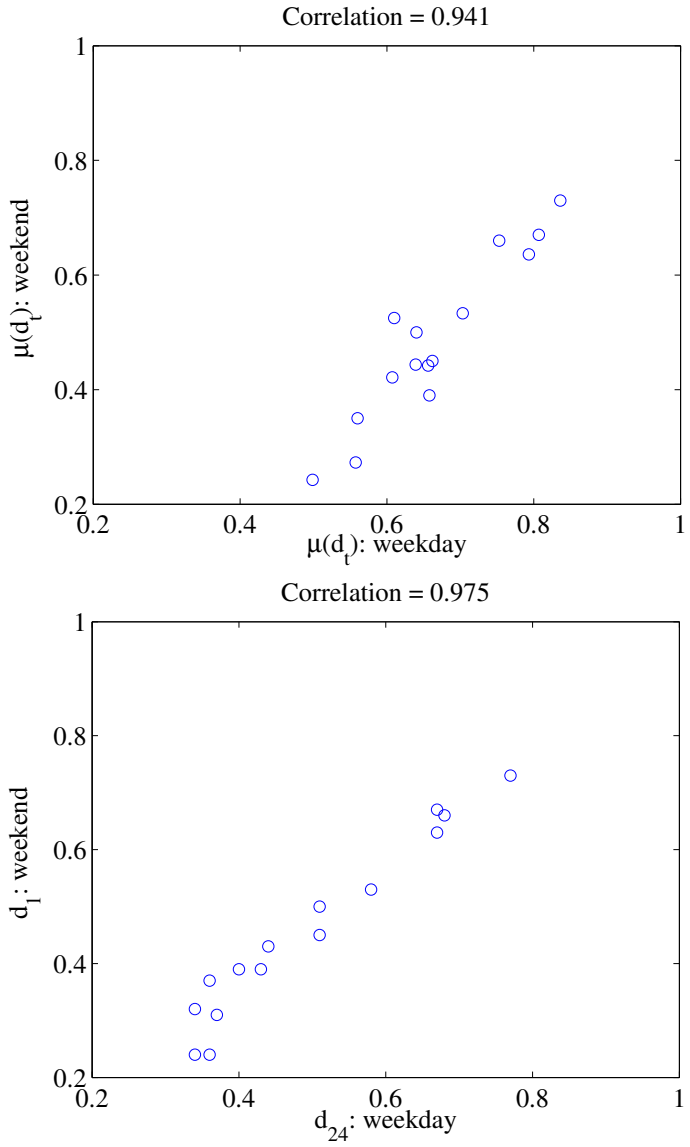


Figure 3.11 Diversity factor correlation of weekday and weekend; left: the averaged diversity factor correlation; right: last hour of weekday and the first hour of weekend

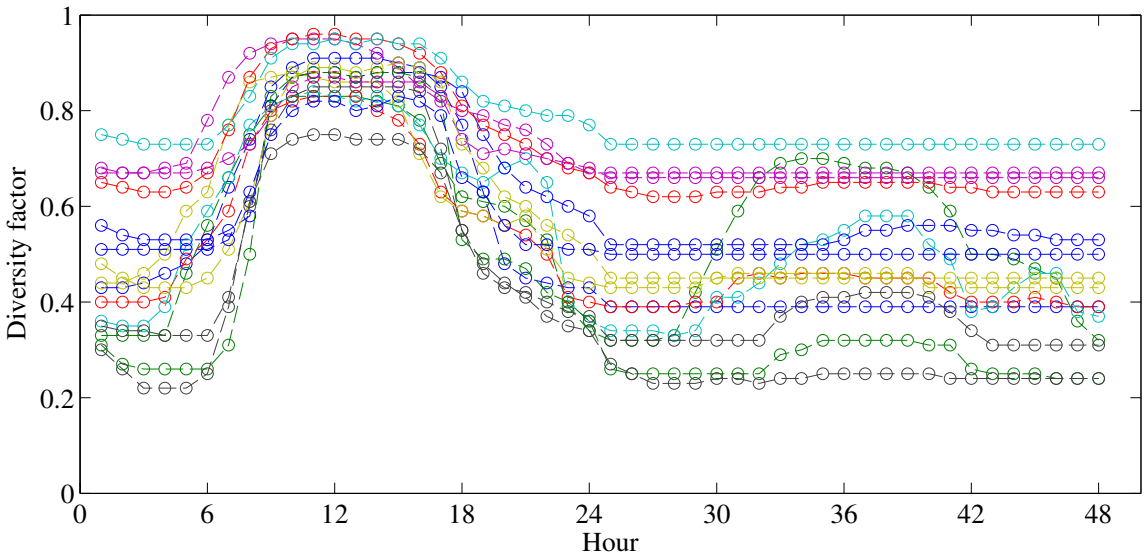


Figure 3.12 Combined diversity factor of weekday and weekend

3.6 An example of model form uncertainty quantification

Modeling solar diffuse irradiation on tilted surfaces is taken as an example to demonstrate the method of model form UQ. Model form UQ based on physical observations can be summarized by six steps as follows:

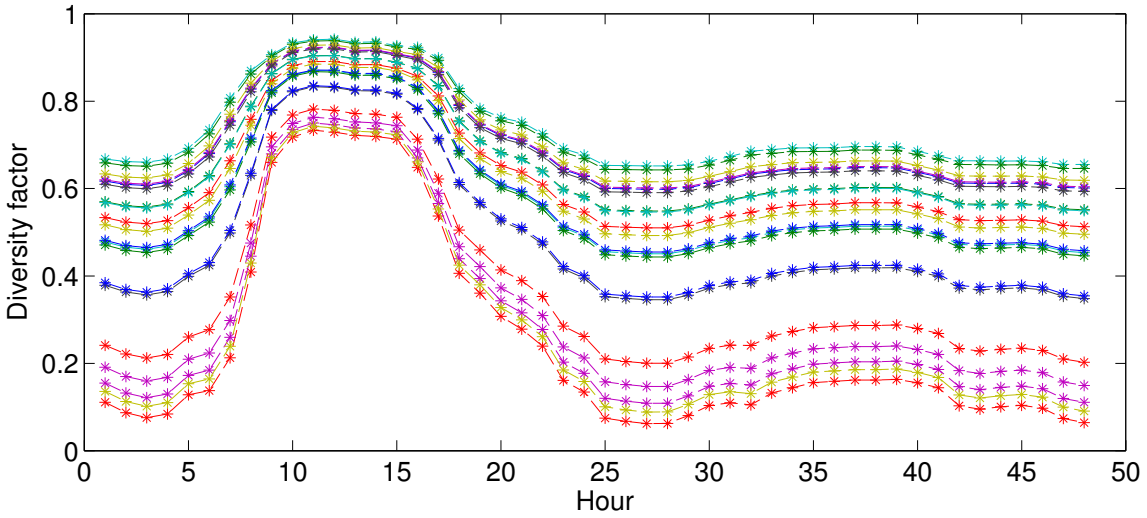


Figure 3.13 20 random samples from multivariate normal distribution with the covariance matrix of $corr(x_i, x_{i+1}) = 1, i = 1,...,23$

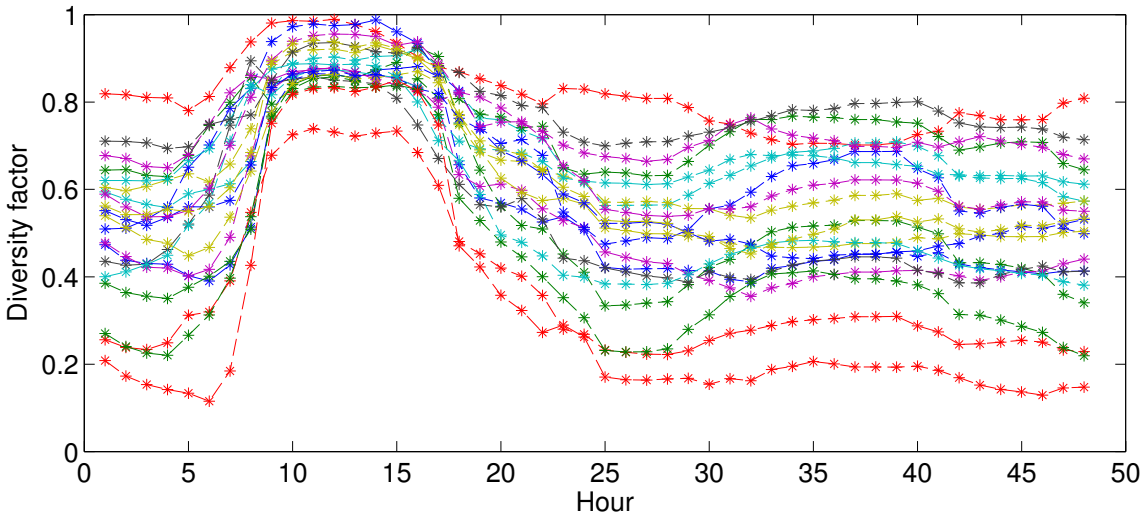


Figure 3.14 20 random samples from multivariate normal distribution using tridiagonal covariance matrix

Step 1: Specify the computational model of interest

Step 2: Obtain physical observations and assess measurement errors

Step 3: Provide statistical evidence for model inadequacy

Step 4: Develop a statistical model from the training data

Step 5: Assess the predictive capability of this statistical model using validation data

Step 6: Update the original computational model and quantify the uncertainties in its predictions

We demonstrate the six steps in the uncertainty quantification of Perez model in the next section.

3.6.1 Solar diffuse irradiation

In building energy models, solar irradiation appears in the boundary condition of external building surfaces. In case of transparent external surfaces it also appears in the boundary conditions of internal surfaces. In rare cases, solar irradiation is directly measured for the orientation and tilt of every surface that appears in a model. In most current building energy studies this is not the case, hence it is derived from other directly

measured quantities and sky condition parameters. The global irradiation on tilted surfaces is calculated by the summation of three components as follows:

$$I_g^s = I_{bn}\cos\alpha + I_d^s + I_r^s \tag{3.8}$$

where I_g^s is the global solar irradiation on a tilted surface with tilt angle S, I_{bn} the direct normal irradiation, α solar incident angle on the surface, and I_d^s diffuse solar irradiation on the tilted surface, and I_r^s the ground reflected irradiation.

Different models have been developed to derive solar irradiation on surfaces with any tilt angle and orientation from data for horizontal surfaces. Among the three components, calculating the direct irradiation is purely geometric and thus straightforward and identical among the models. With respect to the ground reflected irradiation, most studies adopt the isotropic assumption with which the ground reflected irradiation is estimated from the following equation:

$$I_r^s = \rho\frac{1-\cos S}{2} \tag{3.9}$$

where ρ is the ground albedo. The main difference between the models lies in the way the sky diffuse irradiation component is modeled. The sky diffuse irradiation model started from the simple isotropic sky model assumptions (Liu and Jordan 1961), and gradually transformed into anisotropic models advanced by Gueymard (1987), Perez et al. (1990), and Muneer (2004). An extensive literature review of such models and their comparison can be found in (Noorian, Moradi, and Kamali 2008; Gueymard 1987). Among the different models, the most notable one was developed by Perez et al. (1990), as evidenced by its wide application in solar engineering (Yang, Lu, and Zhou 2007) and building energy simulation software such as EnergyPlus (2012). We refer to this particular model as the *Perez model* in this paper. Although the core of the Perez model

pertains to the modeling of sky diffuse component, the model performance is commonly evaluated on the global irradiation on tilted surfaces (Perez et al 1990, Gueymard 2009). Thus, the ground reflected component calculated from equation (3.8) is regarded as an integral part of the broad Perez model. To be consistent with previous studies including Perez's own work, we quantify the uncertainty of the Perez model in predicting the global solar irradiation on tilted surfaces.

3.6.2 Perez model description

This section offers a brief overview of the Perez model (Perez et al. 1990) in terms of modeling the sky diffuse irradiation. It postulates a simplified sky representation, in which the sky hemisphere is composed of a circumsolar disc and horizon band on an isotropic background. Each element has a parametric representation of solar irradiation with multiple coefficients, whose values were obtained through statistical regression analysis. Figure 3.15 shows its input-to-output relationship. The Perez model takes horizontal solar irradiation I_h, direct normal solar irradiation I, solar azimuth angle θ, solar altitude angle φ, surface tilt angle S, and surface azimuth angle ψ as input variables. It calculates diffuse solar irradiation from the sky horizon band $I_{horizon}$, the sky dome I_{dome}, the circumsolar region $I_{circumsolar}$, and total diffuse irradiation from the sky I_{sky}, the latter being the summation of the three components.

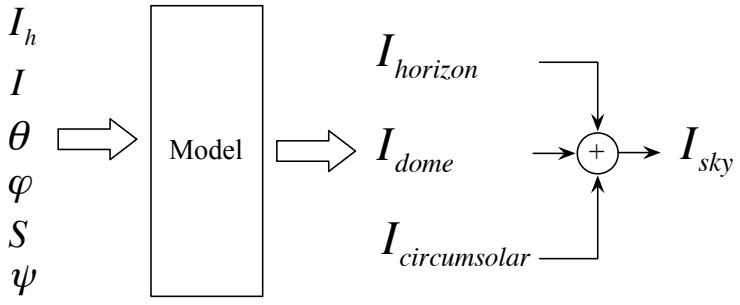


Figure 3.15 Input-to-output relationship of the Perez 1990 sky irradiation model

3.6.3 Experimental data

We obtained detailed measurement data from a station of Solar Radiation Monitoring Laboratory (SRML) in Eugene, Oregon. The data include simultaneous measurements of (1) global solar irradiation on horizontal surfaces, (2) diffuse solar irradiation on horizontal surfaces, (3) direct normal solar irradiation, (4) ground reflected solar irradiation, (5) global solar irradiation on south tilted surfaces at 30°, 45°, and 90°, and (6) global solar irradiation on a north vertical surface. Global solar irradiation (i.e., the sum of direct and diffuse) is measured with the Eppley Precision Spectral Pyranometer (PSP). Diffuse solar irradiation is measured with the shaded Eppley PSE with automatic trackers. Direct normal is measured with Eppley Normal Incident Pyrheliometer (NIP). The ground reflected solar irradiation is measured with the Eppley PSP facing the ground. All devices are yearly calibrated. Table 3.4 provides an overview of the specifications of the instruments.

The data were collected in 2011 at 5 minute intervals. Before we conduct the analysis, we first derive hourly measurements from the raw data. The hourly aggregation reduces the short-term variation of the measurements attributable to small cloud variations and random errors form instrument measurements. Hourly data also matches with the temporal resolution of weather variables used for building simulation such as EnergyPlus (EnergyPlus 2012). We ignored measurements that are less than 50W globally on a horizontal surface because measurements of low solar irradiation are often subject to high measurement error (Reda 2011). Model inputs for solar angles θ and φ are computed according to the ASHRAE Handbook of Fundamentals (ASHRAE 2009b). Uncertainties in the angles are very minimal and are hence ignored.

There is a need to detect systematic measurement errors before performing the UQ. Undoubtedly, undetected systematic measurement errors will contaminate the entire UQ results. In our case, we compare three independent measurements on horizontal

Table 3.4 Specifications of instruments

Instrument	Specifications
Eppley PSP	*Cosine Response:* ±1% from normalization 0-70° zenith angle; ±3% 70-80° zenith angle. *Accuracy:* The absolute accuracy of calibration is about ±3-4%. The relative accuracy of calibration is about ±2%.
Eppley NIP	*Accuracy:* The absolute accuracy of calibration is about ±2%. The relative accuracy of calibration is about ±1%.

surfaces to estimate the quality of the measurements. In principle, the following equation holds:

$$I_g^h = I_{bn}\cos\alpha + I_d^h \tag{3.10}$$

where I_g^h is the global solar irradiation on horizontal surface, I_d^h diffuse solar irradiation on horizontal surface. For horizontal surfaces, every element in equation (3.10) except α, which is computed with minimum uncertainty, is directly measured. Figure 3.16 compares the two sides of equation (3.10) and depicts the comparisons and linear regression analysis. It shows that direct measurements of global horizontal irradiance match well with the calculations from the beam and diffuse horizontal irradiance components. As suggested by the regression equation, the average difference between the directly measured horizontal global irradiation and that calculated by diffuse and direct components is 5.13 W/m^2, which equals 1.5% of the average global irradiation. Therefore, systematic errors in the measurements are considered small enough to ignore. The major measurement errors are attributed to random errors.

3.6.4 Statistical evidence for model inadequacy

We measure the discrepancies between the model predictions and measurements by the two following statistical indicators: Mean Bias Error (MBE) and Root Mean Square Error (RMSE).

$$MBE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i), \tag{3.11}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3.12}$$

where y_i is the i^{th}measurement, $\hat{y}_i$ the i^{th}model prediction, and N the total number of observations.

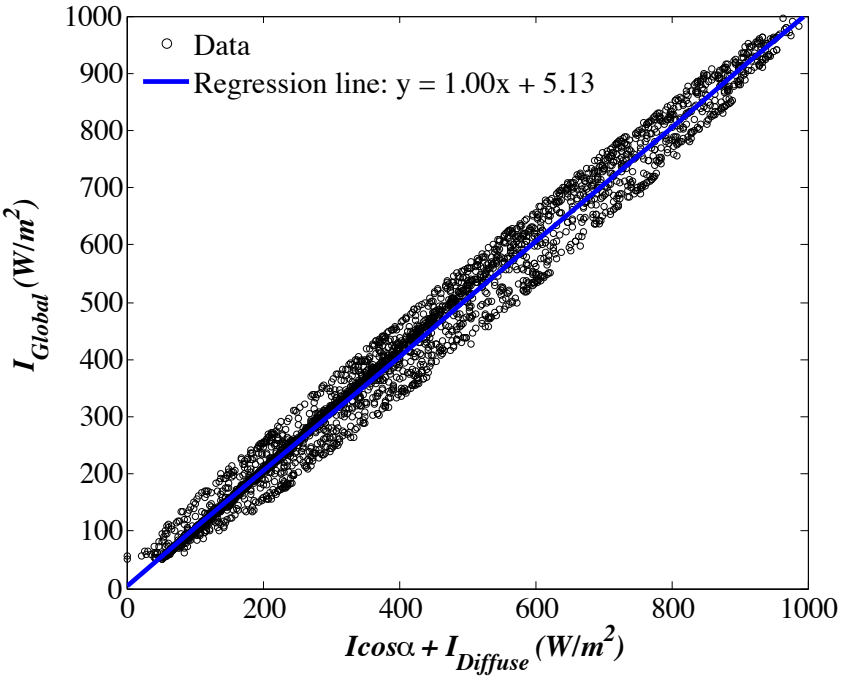


Figure 3.16 Direct measured and derived global horizontal irradiation from two components

We present in Table 3.5 the MBE and RMSE of the global solar irradiation on four different surfaces indicated by orientation and tilt angle, in their original units as well as the percentage to the averaged measurements. It clearly shows that the Perez model over-predicts in all four cases, especially for vertical surfaces. Figure 3.17 compares measured global solar irradiation with Perez predictions on south vertical and north vertical surfaces. It shows that a number of model predictions are higher than measurements. MBE observed on south and north vertical surfaces are $-32\mathrm{W/m^2}$ and $-15\mathrm{W/m^2}$, respectively. The results also show considerable RMSE on the two vertical surfaces, i.e., 22% and 37% for south and north surfaces. Such discrepancies need to be quantified statistically to improve the prediction of the Perez model.

3.6.5 Model form uncertainty quantification

In order to quantify the model uncertainty, we need to build a statistical adjustment to the prediction discrepancies from the Perez model using the Eugene data. The candidate parameters to be used in the adjustment model are the 10 intermediate parameters of the Perez model. They are solar zenith angle, solar azimuth angle, sky brightness factor, sky clearness factor, direct normal irradiation, diffuse horizontal irradiation, global horizontal irradiation, surface tilt angle, surface azimuth angle, and solar incident angle. They are screened by using scatter plots. We plot the prediction discrepancies against the candidate parameters, and check whether the plots display

Table 3.5 Average annual errors of the Perez model in the calculation of the global solar irradiation on four tilted surfaces

	30° South	**45° South**	**90° South**	**90° North**
Average (W/m^2)	396	391	210	67
MBE (W/m^2)	-12	-5	-32	-15
MBE (%)	-3%	-1%	-15%	-22%
RMSE (W/m^2)	47	41	47	25
RMSE (%)	12%	10%	22%	37%

significant patterns. Among the 10 such plots, only four (solar azimuth θ, sky brightness factor Δ, direct normal solar irradiation I, and surface tilt angle S) show some systematic patterns. The corresponding scatter plots are given in Figure 3.18. The plot for the solar azimuth shows a decreasing trend. For sky brightness factor and direct normal solar irradiation, we observe a funnel shape in the plots. This indicates that the variance of $diff$ decreases with Δ and increase with I. The surface tilt angle S is known to be an important parameter in previous works, including Perez (1990). The trend line in this plot shows a negative slope.

The whole year's data are divided into two disjoint parts: the *training* part and the *validation* part. The latter consists of four half-months' observations in January, March, June, and September. Both datasets cover a reasonably complete range of solar angles and sky conditions, so that any potential bias or extrapolation errors can be mitigated.

We use a two-phase regression to construct the adjustment model. Because it has to meet the zero discrepancy constraint on horizontal surfaces (when S=0), we consider a polynomial regression model (Wu and Hamada 2009) in S without constant term. Only the linear and quadratic effects are found to be significant. Therefore, we obtain the

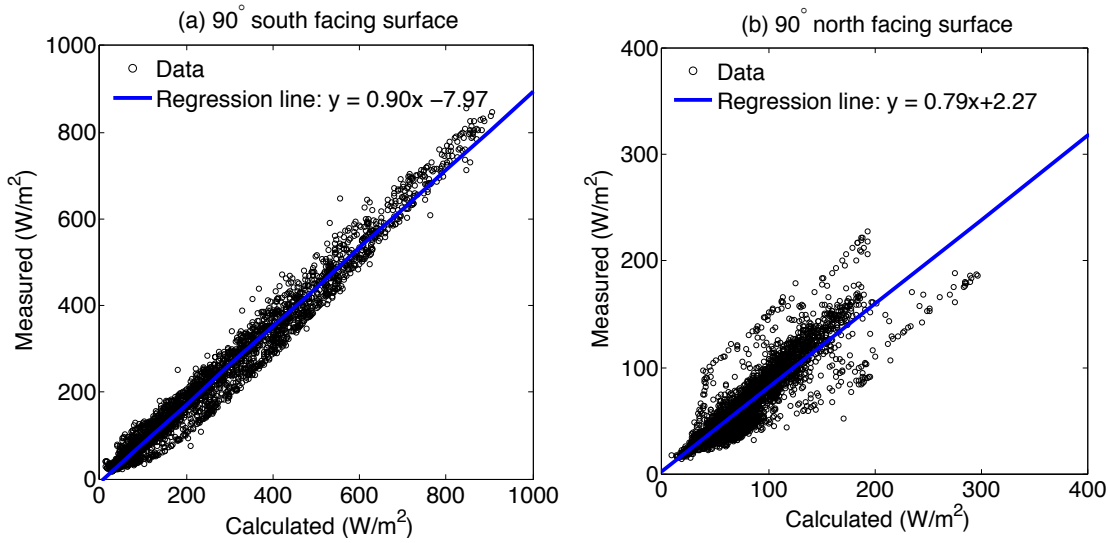


Figure 3.17 Measured and predicted global irradiation on two vertical surfaces

following model for the first phase:

$$diff(\bullet) = f_1 S + f_2 S^2 + \varepsilon, \tag{3.13}$$

where $diff(\bullet)$ is a function mapping inputs into an output.

In the second phase, we fit the coefficients f_1 and f_2 as linear functions of the other three parameters Δ, θ and I. By using a stepwise regression to select parameters, we obtain the following model

$$diff(\bullet) = (c_1 I + c_2\theta + c_3\Delta)S + (c_4\theta + c_5\Delta)S^2 + \varepsilon. \tag{3.14}$$

All the coefficients in Equation (3.14) are significant with p values $<10^{-13}$. Although the

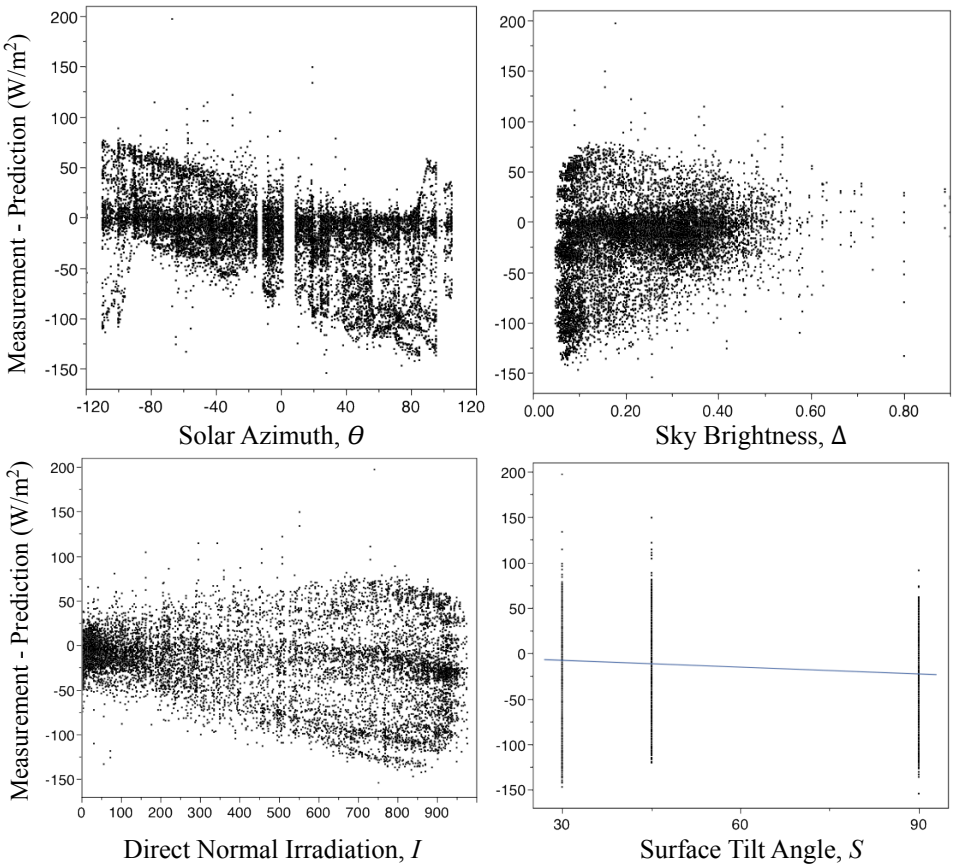


Figure 3.18 Plots of model prediction discrepancies against solar azimuth, sky brightness, direct normal solar irradiation, and surface tilt angle

R^2 value is only 0.504, it is acceptable because there are various sources of measurement errors present in the data. Ordinary residual diagnostics were performed to verify the goodness of fit of the model. The residual plot shows a random pattern and the QQ plot almost forms a 45-degree line. So the fitted linear model appears to be quite good. We have now found a reasonably good adjustment for the Perez model for the Eugene data. We call the sum of the Perez prediction and the fitted regression model in Equation (3.14) the *modified Perez* model.

3.6.6 Validating model form uncertainty quantification

In this section we validate the $diff(\bullet)$ model developed above on the validation dataset, which consists of four half-months' observations in December, March, June, and September. Table 3.6 compares the modified Perez model to the original Perez in terms of MBE and RMSE. The results show a substantial improvement in predicting global solar irradiation on tilted surfaces. The new model reduces the MBE by more than 50% for every test surface. For example, for south vertical surface, it decreases from -33 W/m^2 to -11 W/m^2, and on north vertical surface, it reduces from -15 W/m^2 to 8 W/m^2. In terms of RMSE, the modified Perez model also shows a significant improvement. Particularly for the south vertical surface, the RMSE is reduced by over 40% from 46 W/m^2 to 28 W/m^2. Moreoever, the modified model does not consistantly overpredict. From the energy balance perspective, the modified model promises a more reliable prediction as it avoids significant amounts of overpredictions of solar irradiation on building envelops stemming from Perez model inadequacy. Although the variable selection results indicate that surface azimuth is an insignificant factor of the Perez model discrepancy, we will investigate this result by collecting data on east- and west-facing surfaces in a future study.

3.6.7 Prediction of the modified Perez model

In Table 3.6, the two-phase regression model $diff(\cdot)$ shows its capability of correcting the bias in predicting the mean of solar irradiation on tilted surfaces. New predictions are then simply obtained by the addition of original computational model outcomes from Perez, i.e, $\eta(\boldsymbol{u})$ and that from the adjustment regression model, i.e, $diff(\boldsymbol{u},\boldsymbol{v})$. We can also construct the confidence intervals to estimate the uncertainties in new predictions from the linear regression model (Seber and Lee 2012).

We take two days (June 1 and 2, 2011) from the validation dataset to illustrate the prediction capability of $diff(\cdot)$ and the UQ for the 90° south-facing surface at each hour. Each day contains ten hours of values when the surface is exposed to the sun. We first compare the physical observations y_{obs}, with Perez model predictions y_{Perez}, and illustrate the results in Figure 3.19(a). This figure shows that Perez model yields higher predictions at 18 out of the 20 points than the measurements whereas the maximum model discrepancy appears at around 12:00. The differences, i.e., $y_{obs} - y_{Perez}$, will then become the physical observations on which $diff(\cdot)$ will predict. Figure 3.19 (b) illustrates $y_{obs} - y_{Perez}$, predictions of $diff(\cdot)$, and a 95% confidence interval. It shows that the 95% confidence interval well covers the hourly variation of $y_{obs} - y_{Perez}$. Therefore, we will use $diff(\cdot)$ to quantify the prediction uncertainties given by the original Perez model.

Table 3.6 Model validation statistical results

	30° South		45° South		90° South		90° North	
	Perez 1990	Modified Perez	Perez 1990	Modified Perez	Perez 1990	Modified Perez	Perez 1990	Modified Perez
Average (W/m^2)	385		382		210		68	
MBE (W/m^2)	-14	-7	-7	3	-33	-11	-15	8
MBE (%)	-4%	-2%	-2%	1%	-16%	-5%	-21%	11%
RMSE (W/m^2)	45	31	39	30	46	28	22	19
RMSE (%)	12%	8%	10%	8%	22%	14%	33%	28%

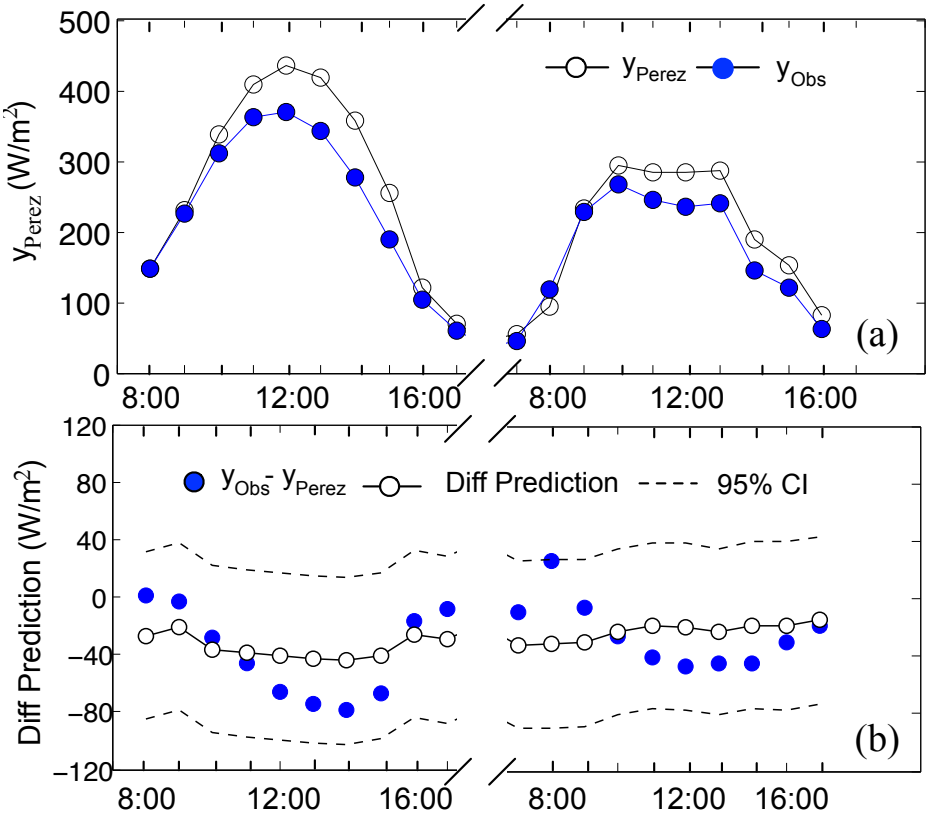


Figure 3.19 Results on a south-facing vertical surface at hourly intervals over two days (June 1 and 2, 2011): (a) Perez predictions, y_{Perez}, and physical observations, y_{obs}, and (b) prediction of $diff(\cdot)$ and 95% confidence interval

Uncertainties that are indicated by the confidence interval come from two major sources: the random errors in the measurements and the remaining model bias not captured by $diff(\cdot)$. The model form uncertainties in the predictions are thus reducible if measurement errors decrease or if more experimental data are added. The remaining model bias can also reduce if additional variables (as explained earlier, these variables will be part of $\boldsymbol{v}$) are detected to be significant and are then added to the $diff(\cdot)$ model in a sufficient manner.

3.7 An uncertainty quantification repository

In the previous sections, we introduced the methods of UQ in parameters of different types, and in model forms, using different examples. For the whole building energy models, a comprehensive UQ is unarguable the most challenging and critical part of the whole UQ project, yet has not received enough attention as it deserves. The previous examples are the UQ cases the author of the thesis has been closely involved. More examples on UQ of other parameters and more forms are reported in our previous papers (Sun, Su, et al. 2014; Sun, Heo, et al. 2014; Sun and Augenbroe 2014; Lee et al. 2013a; Augenbroe et al. 2013) and reports (Augenbroe 2012). One major objective of this research is to perform an exhaustive UQ across five spatial scales and in two types as introduced in Section 3.3. Table 3.7 summarizes the types and sources of uncertainty. Results of all UQ are stored in a UQ repository in XML format. The repository can be accessed using an interface developed in Microsoft Excel. Detailed introduction about the UQ repository is referred to our previous paper (Lee et al. 2013a). This UQ repository is to be used for the following uncertainty analysis in the thesis.

				IDF INPUTS							
			Instance Name	IsQuantity	IsUncertain	DistributionType	Parameter 1	Parameter 2	Parameter 3	Parameter 4	Error Check
				Is this a Number?	Is the number uncertain?	What kind of distribution defines the uncertainty?	Lower Bound, Mean	Upper Bound, St. Dev.	Minimum	Maximum	Are the parameters logical?
BuildingUQ	ConvCoeff	ConvCoeff InternalUQ	IncludeInteriorHcforUQ	FALSE	FALSE						
			AlphaHnWall	TRUE	TRUE	UniformAbsolute	0	1	-999	-999	1
			NHnWall	TRUE	TRUE	UniformAbsolute	0	1	-999	-999	1
			AlphaHnFloor	TRUE	TRUE	UniformAbsolute	0	1	-999	-999	1
			NHnFloor	TRUE	TRUE	UniformAbsolute	0	1	-999	-999	1
			AlphaHnCeiling	TRUE	TRUE	UniformAbsolute	0	1	-999	-999	1
			NHnCeiling	TRUE	TRUE	UniformAbsolute	0	1	-999	-999	1
		ConvCoeff ExternalUQ	IncludeExteriorHcforUQ	FALSE	FALSE						
			AHext	TRUE	TRUE	UniformAbsolute	0	1	-999	-999	1
			BHext	TRUE	TRUE	UniformAbsolute	0	1	-999	-999	1
	Infiltration LowRise		IncludeInfiltrationUQforLowRise	FALSE	FALSE						
			BuildingHeight	TRUE	FALSE						
			BuildingWidth	TRUE	FALSE						
			BuildingLength	TRUE	FALSE						
			ELAPerExteriorArea	TRUE	TRUE	LogNormalAbsolute	-1.6551	0.8767	-999	-999	1
			RFracLeakAreaFloorToCeiling	TRUE	TRUE	UniformAbsolute	0	1	0	-999	1
			XDiffCeilingToFloorLeakArea	TRUE	TRUE	UniformAbsolute	0	1	-1	1	1

Figure 3.20 Portion of Excel interface for accessing XML files in a UQ repository

Table 3.7 Types and sources of uncertainties in a UQ repository

Uncertainty sources	Uncertainty types	
	Model Form	**Parameter**
1. Meteorology		
Weather	✔	
2. Urban		
Local wind speed	✔	✔
Urban heat island effect	✔	✔
Wind pressure coefficient	✔	✔
Ground reflectance		✔
Diffuse solar irradiation on tilt surfaces	✔	
3. Building Shell		
Interior convective heat transfer coefficients: wall, ceiling, and floor		✔
Exterior convective heat transfer coefficients		✔
Infiltration	✔	✔
Internal furniture		✔
Thermal bridge	✔	✔
Single-side natural ventilation	✔	✔
Room temperature stratification		✔
Material properties		✔
4. Systems		
HVAC equipment efficiency due to manufacture variability including chiller, fan, pump, etc.		✔
Controls of HVAC operation	✔	✔
5. Operation		
Lighting and plug load density		✔
Lighting and plug load schedules		✔
Occupant density		✔

```
<AlphaHnWall>
    <IDFInput>
        <IsQuantity>true</IsQuantity>
        <IsUncertain>true</IsUncertain>
        <DistributionType>UniformAbsolute</DistributionType>
        <Parameter1>0</Parameter1>
        <Parameter2>1</Parameter2>
        <Parameter3>-999</Parameter3>
        <Parameter4>-999</Parameter4>
    </IDFInput>
</AlphaHnWall>
<NHnWall>
    <IDFInput>
        <IsQuantity>true</IsQuantity>
        <IsUncertain>true</IsUncertain>
        <DistributionType>UniformAbsolute</DistributionType>
        <Parameter1>0</Parameter1>
        <Parameter2>1</Parameter2>
        <Parameter3>-999</Parameter3>
        <Parameter4>-999</Parameter4>
    </IDFInput>
</NHnWall>
<AlphaHnFloor>
    <IDFInput>
        <IsQuantity>true</IsQuantity>
        <IsUncertain>true</IsUncertain>
        <DistributionType>UniformAbsolute</DistributionType>
        <Parameter1>0</Parameter1>
        <Parameter2>1</Parameter2>
        <Parameter3>-999</Parameter3>
        <Parameter4>-999</Parameter4>
    </IDFInput>
</AlphaHnFloor>
```

Figure 3.21 A example of XML schema for three parameters in a UQ repository

3.8 Uncertainty analysis

Sample-based methods such as Monte Carlo based simulations, are widely used in UA with application examples across many engineering domains (Allaire 2009; Helton 2009; Helton and Davis 2003; Janssen 2013). Our UA is based on Monte Carlo simulation. The concept of Monte Carlo UA is illustrated in Figure 3.22 using two uncertainty inputs $x_1$and x_2. In general, denote the model input by $\boldsymbol{x} = (x_1, \dots, x_p)$, and the model output $y = f(\boldsymbol{x})$. The uncertainty in each parameter $x_j,\ j = 1, \dots, p$, is quantified by a probability distribution. We generate n samples consistent with the distributions of parameters. The sampling process is also named as design of experiments. The samples are then executed in the model $y = f(\boldsymbol{x})$ to generate the corresponding outputs of y. The empirical CDF of y is then drawn from the simulated outputs. As the number of samples increases, the empirical CDF converges to the true CDF of y with the rate of $\sqrt{n}$. Thus, the performance of Monte Carlo simulation is independent of the number of uncertain parameters p, which is the major advantage when p is very large, e.g., 100s or 1000s.

Since the samples carry the probability information, how to obtain the samples to well explore the p-dimensional space is essential. The method we used to sample inputs for the simulation is Latin hypercube sampling (LHS), which was first proposed by Mckay, Beckman, and Conover (1979). Instead of random sampling, LHS provides robust results in relatively small samples (Saltelli, Tarantola, and Campolongo 2000). Another more attractive property of LHS will be discussed in the following SA section. LHS operates as follows: For any uncertainty parameter $x_j, j = 1, \ldots, p$, the range of the parameter is divided into n equal probability intervals. Randomly sample one value from each interval. Then take an arbitrary permutation of x_j as the jth column of the sample matrix representing the sample values of parameter j. Combining the columns of all p parameters, we have the Latin hypercube samples, where the sample matrix has n rows (samples) and p columns (parameters). Figure 3.23 uses two uncertain parameters, $X_1 \sim N(10,1^2)$ and $X_2 \sim U(1,10)$ to illustrate the LHS design process for 10 samples. Detailed technical description and other types of design of experiments is referred to (Santner, Williams, and Notz 2003; Wu and Hamada 2009).

3.9 Sensitivity analysis

Sensitivity analysis (SA), and probabilistic SA in particular, is concerned with how the total output uncertainty can be attributed to uncertainties in individual inputs or groups of inputs. SA ranks the importance of input parameters based on their influence

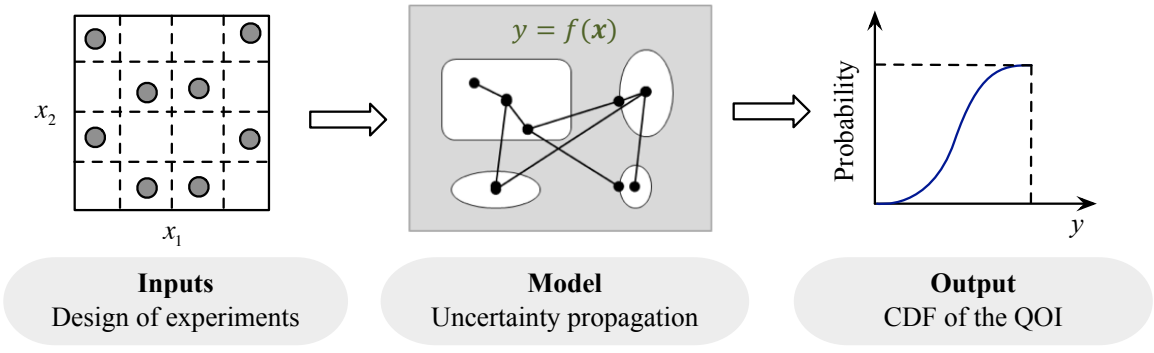


Figure 3.22 Monte Carlo simulation for uncertainty analysis

on the uncertainties of model outcomes. The role of SA can be multiple. In some cases SA is used to single out the most important sources of uncertainty to start a next step in which these parameters will be better quantified. In other cases SA is used to identify the sources of risk with the aim of reducing the uncertainties by implementing better quality assurance or guarantee reduced uncertainty in performance contracts.

UA and SA are closely related, but SA is more difficult than UA and may need a special design of computer experiments (samples) to construct the ranking of parameters according to their influence. When the number of uncertain parameters gets large, which is the case for most buildings, brute-force probabilistic SA becomes prohibitively time-consuming because it requires tens of thousands of simulations to enable the calculation of the global sensitivity measures. Although these SA methods have their generic use in engineering models (Saltelli et al. 2008), they tend to be inefficient and sometimes over-

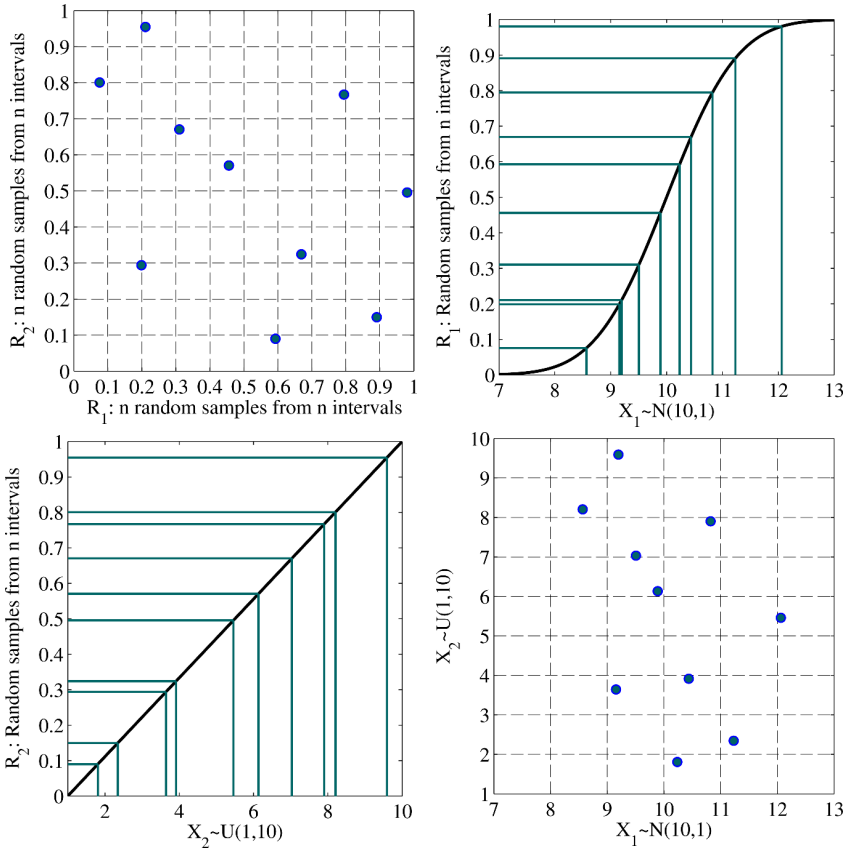


Figure 3.23 Latin hypercube design for two uncertain parameters

engineered for a particular application. Here, we use a regression-based SA method that can better fit the needs of building energy simulation than the generic ones.

The SA method is accomplished as a post-processor, sharing the same LHS design and simulation runs with the preceding UA. The SA method entails two steps. First we use parameter screening to remove insignificant parameters. The left-out parameters then enter the second step, in which the sensitivity indices (SI's) are computed based on variance decomposition.

We consider the main effects of parameters with associated uncertainties by the following linear regression model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip,} i = 1,\ldots,n, \tag{3.15}$$

where y_i is the model output, x_{ij}'s, $j = 1, \ldots, p$, are the model inputs, β_j's, $j = 0, \ldots, p$, are the regression coefficients.

Parameter screening is necessary before we compute the SI due to the following reasons. The accuracy of sensitivity measurements will decrease if there exist high correlations between significant and insignificant parameters in samples. This becomes evident when limited runs, e.g., 1000, are used to simultaneously analyze a large number of uncertainty parameters, e.g., 100s. Thus, the insignificant parameters need to be removed before computing the SI. The theoretical foundation to use parameter screening relates to the effect sparsity principle in statistics, i.e., only a few parameters are significant among many candidates (Wu and Hamada 2009).

We use lasso (Tibshirani 1996) as the screening method. The procedure is briefly introduced here. The lasso estimates of the coefficients β_j's in equation (1) are given by

$$\hat{\beta}_{lasso} = argmin_{\beta}\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}, \tag{3.16}$$

where λ is the tuning parameter that controls how many parameters are selected in the model. Larger λ generates a smaller model. Compared with the ordinary least squares (OLS) estimates,

$$\hat{\beta}_{OLS} = argmin_{\beta}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2, \tag{3.17}$$

lasso penalizes large β_j's, and shrinks the estimates of them. All lasso estimates of the β_j's are hence uniformly less than OLS estimates. Importantly, some of them that have small OLS estimates are directly shrunk to be zero because of the property of L_1 penalty (the last term in (2)). Thus, it is very effective in parameter screening. The performance has been proven in the statistical literature. Efron et al. (2004) gave a more intuitive explanation of lasso as the screening method: lasso keeps increasing the coefficients in the "best" direction, and always chooses the most correlated parameter to enter the model. See also Hastie et al. (2009).

We point out one important property of LHS when only a few of the parameters are significant. Note that n values uniformly cover all n intervals for any parameter, which holds independent of other parameters. No matter which parameters are significant, the design after screening is still composed of Latin hypercube samples with all the desirable features of LHS.

After p_0 significant parameters are identified, we turn to the second step to calculate SI's based on ANalysis Of VAriance (ANOVA). In ANOVA, the total variation associated with the outputs is measured by total sum of squares (SST). SST can be split

into two components: regression sum of squares (SSR), i.e. the variation in the outputs explained by the model, and error sum of squares (SSE), i.e. the variation in the outputs not explained by the model. The ratio of SSR to SST is the determinant of coefficient, i.e., R^2, indicating the goodness of fit using linear model (1). Further decomposing SSR on each parameter leads to the variation in the outputs explained by parameter j. Finally, sensitivity index (SI) for parameter j is computed by Equation (4).

$$SI_j = \frac{SS_j}{SST} \times 100\%. \tag{3.18}$$

SI_j measures the percentage of the variation in model response y that is attributable to the uncertainty in parameter j. The sum of SI's over all parameters is less than or equal to 100%.

Sometimes, parameters can be classified into several groups. The group SA is more interesting and easier to comprehend than the SA for individual parameters. If so, group SA can be done after SI's of all significant parameters have been computed. The group SI is equal to the total SI's of the parameters that belongs to this group. It is also possible to analyze discrete parameters, e.g., different weather years, in our model using dummy variables (Faraway 2004).

CHAPTER 4 PREDICTION VERIFICATION

4.1 Introduction

Making predictions over some quantity or event in the future is one of the major human desires, primarily for decision making. In the building domain, predictions made at the design stage could be energy performance, air quality, acoustical and lighting environment, and rent premium, etc. The predictions could be obtained from expert knowledge, a simple calculation, or a dynamic computer simulation. Furthermore, the predictions can be delivered as a real-valued number (i.e. point or deterministic prediction) or as a probability distribution. For all these situations, it is always appealing to assess how well the predictions come true when the predicted events or quantities materialize. It is also very common that different models offer distinct predictions for the same modeling problem, or a new modeling system is developed to replace its old version. Accordingly, one inevitably wants to discern the best from a set of competing ones, or the improvement of a new version compared to its old version. This chapter deals with the topic of prediction verification, aiming to support model prediction assessment in a rigorous and logical way.

4.2 A theoretical framework for prediction verification

We develop a theoretical framework for prediction verification derived from (Gneiting 2011; Gneiting and Katzfuss 2014). The framework consists of the following components:

- A *prediction-observation* (PO) domain, $\Omega = \mathrm{P} \times \mathrm{O}$, the Cartesian product of the domain $\mathrm{P} \subseteq \mathbb{R}^d$ with $\mathrm{O} \subseteq \mathbb{R}^d$.
- A *scoring function* $\mathrm{S}: \Omega \to [0, \infty)$, where $\mathrm{S}(x, y)$ represents the score or loss when the prediction x is issued and the observation y materializes.

- A *weighting function* that aggregates the scores or losses over multiple cases to derive a global measure on which the model is assessed.

Let us take building energy use intensity (BEUI) as the quantity of interest to illustrate this framework. Suppose a modeler provides a point (i.e. deterministic) prediction at the design stage, denoted by x; and annual energy consumption is measured for a given year, denoted by y. We define a scoring function S as follows:

$$\mathrm{S}(x,y) = |x - y|, \tag{4.1}$$

which computes the absolute difference between predicted and observed value. The score will then represent the skill of the modeler on this particular building evaluated at a given year. Now, suppose we have a number of buildings with predictions and observations coming from the same modeler, we want to have a summary measure for this modeler. A weighting function is then applied to aggregate the scores. Let n denote the number of buildings under consideration, and each building has m observations, so we have $m \times n$ evaluation points. A possible weighting function may be the simple average of all scores, i.e.:

$$\bar{\mathrm{S}} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{S}_{ij}. \tag{4.2}$$

Obviously, the scoring function is a critical component in the framework. It is relatively straightforward to define scoring functions for point predictions. In our framework, the scoring function is negatively orientated, i.e. the smaller, the better. There are many functions that can be used as the scoring functions as long as they have the following property: $\mathrm{S}(x,y) \geq \mathrm{S}(y,y)$ with equality if and only if $x = y$. A scoring function

with this property is called to be strictly proper (Gneiting and Raftery 2007). We shall introduce such scoring functions and discuss their properties for point prediction verification as well as for probabilistic prediction verification. A weighting function offers a compact measure over a variety of verification cases and thus is used to compare competing prediction procedures.

4.3 Point prediction verification

Point prediction is the dominant form for predicting building performance. Some commonly used functions to evaluate model predictions are listed in Table 4.1. Among the four commonly used functions, squared error (SE), absolute error (AE), and absolute percent error (APE) are strictly proper scoring functions. However, the bias error is not strictly proper because the total errors will be evened out by the positive and negative errors.

If we use the simple average as the weighting function, the corresponding aggregated scores are mean squared error (MSE), mean absolute error (MAE), and mean absolute percent error (MAPE), and mean bias error (MBE). Even though these scoring functions are widely used in point prediction evaluation, there is not enough recognition that they may yield conflicting results in the comparison of competing models. To illustrate the idea, let us evaluate three solar irradiation models that compute global solar irradiation on tilted surfaces using horizontal measured diffuse and direct components. The three models are the simple isotropic model, the Perez model (Perez et al. 1990), and another model by Hay and McKay (1988). The measurement data are obtained from the

Table 4.1 Commonly used functions to evaluate building performance simulation

Scoring function	**Name**
$S(x, y) = (x - y)^2$	Squared error (SE)
$S(x, y) = \|x - y\|$	Absolute error (AE)
$S(x, y) = \|(x - y)/y\|, y \neq 0$	Absolut percent error (APE)
$S(x, y) = (x - y)$	Bias error (BE)

National Renewable Energy Laboratory. We compute the hourly AE, APE, and SE on predicting the global solar irradiation on a south-facing vertical surface over 50 non-zero consecutive hours. Figure 4.1 plots the absolute errors over the 50 hours. It shows that AE fluctuates from hour to hour. Figure 4.2 and Figure 4.3 show the absolute percent errors and the squared errors, respectively. Obviously, we do not have an absolute winner whose prediction error is smaller than the others for any given hour.

To evaluate the three models, we calculate the simple average over all evaluation points to derive a summary measure for each model on which the model comparisons are based. Table 4.2 shows the MAE, MAPE, and MSE and the corresponding ranks. The result shows that the model by Hay ranks third using this set of verification data. More interestingly, we rank the isotropic model the best based on MSE, whereas we rank the Perez model the best based on MAE or MAPE.

The example shows that the rankings for competing models may be inconsistence

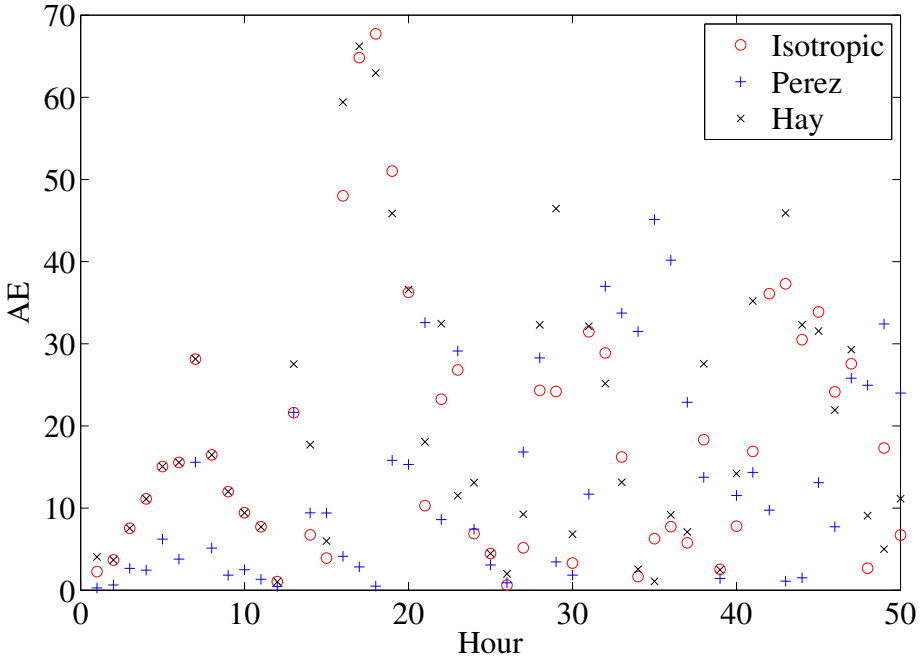


Figure 4.1 Absolut errors (AE) of three solar irradiation models evaluated on the vertical south-facing surface

if different scoring functions are used. In fact, there does not exist a ranking r of any model predictions p_i and p_j both distinct from observations o, such that for all scoring functions,

$$r_i \geq r_j \Leftrightarrow \bar{S}_i \geq \bar{S}_j . \tag{4.3}$$

Hence, the ranking of model predictions should not be interpreted independent of the scoring functions. In other words, different scoring functions will have agreed rankings only if the prediction is perfect (i.e. the one that coincides with observations). However, every scoring function has its own penalty mechanism on the prediction cases deviated from the observations. For example, SE gives larger penalty to the larger prediction errors than the smaller ones whereas the AE is indifferent to the magnitude of errors.

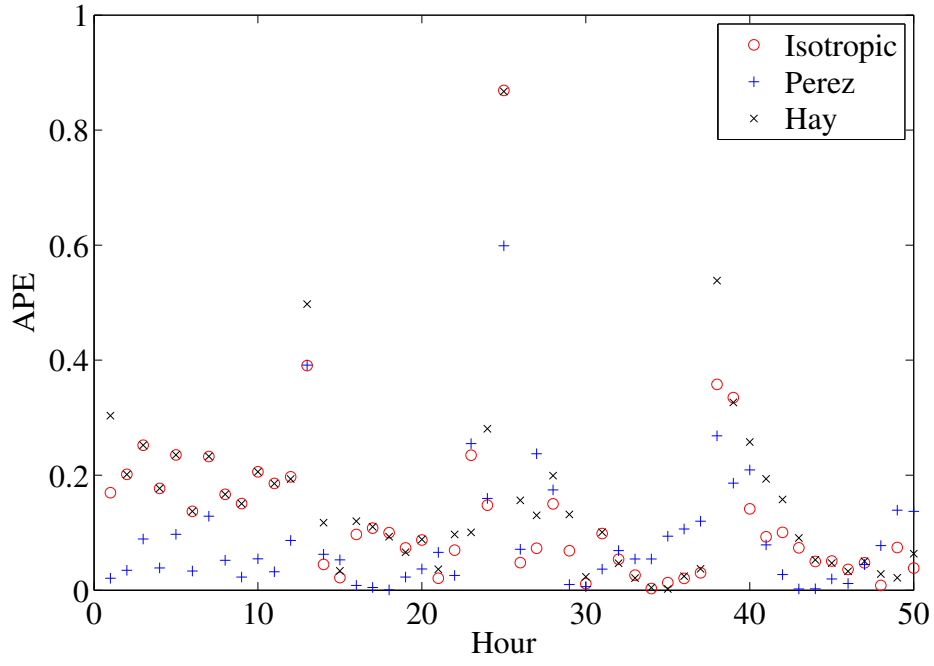


Figure 4.2 Absolut percent errors (APE) of three solar irradiation models evaluated on the vertical south-facing surface

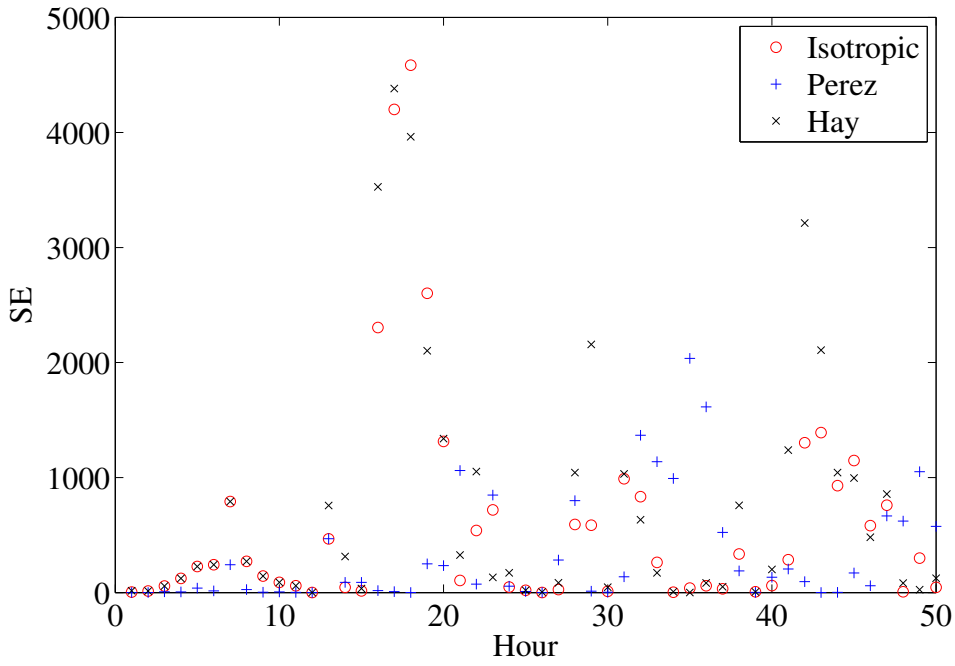


Figure 4.3 Squared errors (SE) of three solar irradiation models evaluated on the vertical south-facing surface

Table 4.2 Comparing three models by different scoring functions

Model	Scoring function			Ranking by		
	MAE	MAPE	MSE	MAE	MAPE	MSE
Isotropic	13.4	0.13	306.8	2	2	1
Perez	12.8	0.09	337.8	1	1	2
Hay	16.7	0.15	461.9	3	3	3

4.4 Probabilistic prediction verification

Probabilistic predictions allow us to express the uncertainty in estimating a quantity of interest that typically pertains to the future. Thus, it enables risk assessment in decision-making. Because of the advantages of probabilistic prediction, the last decade has seen a surge of interest in uncertainty quantification in the building research domain. As most technical issues, such as how to propagate the uncertainty, have been more or less resolved, some fundamental questions start to rise. Just to give a few examples here,

we raise two questions: (1) are probabilistic predictions guaranteed to be better than point predictions in the support of decision-making; and (2) how to assess two different probabilistic predictions from two competing models?

From the prediction form perspective, the probability distributions contain more information than point predictions. In particular, probabilistic predictions provide all possible estimates and the associate likelihoods are conditional on the level of knowledge when the predictions are made. It is a logical reflection of our partial understanding of reality using an imperfect model. Thus, we can conclude that using probability as the form of model prediction enables a more complete representation of our understanding of the reality, and thus is an improvement compared with a point prediction. From an application perspective, however, it is possible that point predictions from a skilled modeler can outperform probabilistic predictions from a novice. Generally, probabilistic predictions are just another form of information. Thus, there is no guarantee that the quality of information from one form of prediction is better than the other. In fact, prediction verification should be performed for both point and probabilistic prediction.

4.4.1 Probabilistic prediction and verification data

Here, we focus on the probabilistic predictions on building performance provided at the design stage. A probabilistic prediction, D, is provided, taking the form of cumulative density function (CDF). For example, D could be the BEUI distribution, which utilizes a certain knowledge and information basis, including the modeler expertise, theories coded in the model, and data specific to this design. Let us imagine that there is an ideal probability distribution, *F,* that uses all information that is currently accessible and permitted. For our example, if the design specifications were to be turned into realizations multiple times and we could simultaneously obtain multiple observations, these observations would approximate the distribution *F*. Of courses, there are no strict replications coming from the same design because, even if we are dealing

with nearly identical buildings, every exemplar is built at a different site with different surroundings. However, if the building energy model is capable to account for the effects of these factors, we can still treat them as nearly replications.

In addition, the building energy performance is traditionally predicted using one whole year, i.e. 8760 hours as the prediction period without specification of the actual time period for the prediction to take effect. Monthly or annual results are obtained from post-processing. Following this tradition, the probability distribution of BEUI also contains the uncertainty due to unknown future weather at the design stage. Hence, observations at different years for a fixed building may also be treated as multiple realizations from the distribution of the weather, conditional on the realized values of the other parameters. However, this should not be confused with measurements at different months or hours. There will be different CDF's for different months, i.e. 12 CDF's representing the distribution of energy use during each month. It makes no sense to construct a CDF using samples from different months and then compare with an empirical monthly CDF constructed from monthly measurements. Suppose we have multiple realizations from which an empirical CDF is constructed. This empirical CDF can be used to assess the predicted CDF, *D,* and certain statistical measures can be computed. In reality, this situation may apply to the residential domain where nearly identical buildings are implemented from one design, but it is rarely the case for commercial buildings. In spite of that, there are many known parameters that significantly influence building energy performance. These parameters occur at different scales and at different stages of the whole building delivery process. It is not likely that we could obtain a reliable empirical CDF from one building design. The task of assessing the CDF prediction appears difficult, perhaps impractical. As it turns out, the challenges posed by the verification data constraints are not insurmountable. There is method that can pool information from different buildings with different design specifications to assess the probabilistic predictions simultaneously discussed below.

4.4.2 Probability integral transform

We introduce our first method for probabilistic prediction verification: the probability integral transform (PIT). This method has been widely used to verify probabilistic weather forecasting (Jolliffe and Stephenson 2003). Arguably, weather prediction is the most mature and successful implementation of probabilistic predictions. There is a long history of forecasting weather in a probabilistic way. For example, there is a 30% chance of raining tomorrow. The result, i.e. the verification data, will be either rain or not rain; thus it seems to be impossible to verify the validity of the 30%. However, this is not insurmountable if we realize that there are many probabilistic predictions issued with one real-valued observation for each. PIT can pool information from different incompatible situations on which the underling model or modeling procedure is assessed. Modeling procedure is defined as a series of actions and guidelines leading to the development of a model. In other words, different models may share one modeling procedure. Modeling procedure can thus be regarded as a population with samples referred to models of different systems.

For a continuous variable, the probability integral transform, Z_F, is simply the

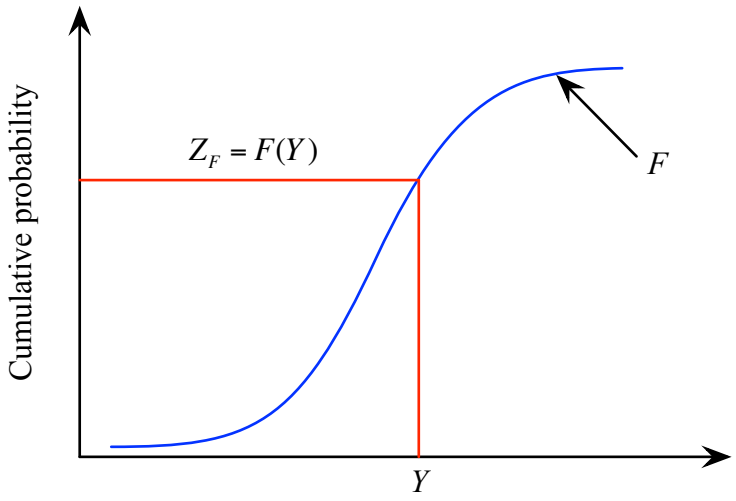


Figure 4.4 Probability integral transform (PIT)

predicted CDF, F, evaluated at an observation point at Y, i.e. $Z_F = F(Y)$. Figure 4.4 depicts the PIT. Recall that Y is a random variable, which, for instance, is the future building energy consumption at its design stage. We want to predict the distribution of Y to make risk-conscious decisions. F is the CDF, representing our prediction of the distribution of Y. If the predicted F is ideal, then Y should follow the distribution F, i.e. $Y \sim F$. If so, Z_F follows a standard uniform distribution.

Additionally, the uniformity can be generalized to any arbitrary CDF. Therefore, we can construct an empirical distribution for Z_F coming from different CDFs evaluated at its corresponding observation point at Y. Figure 4.5 illustrates the idea. N PIT, i.e. $Z_i, i = 1,...,N$ from different predicted CDF and one corresponding real-valued observation $Y_i, i = 1,...,N$ are pooled into one CDF of Z. Under the assumption that

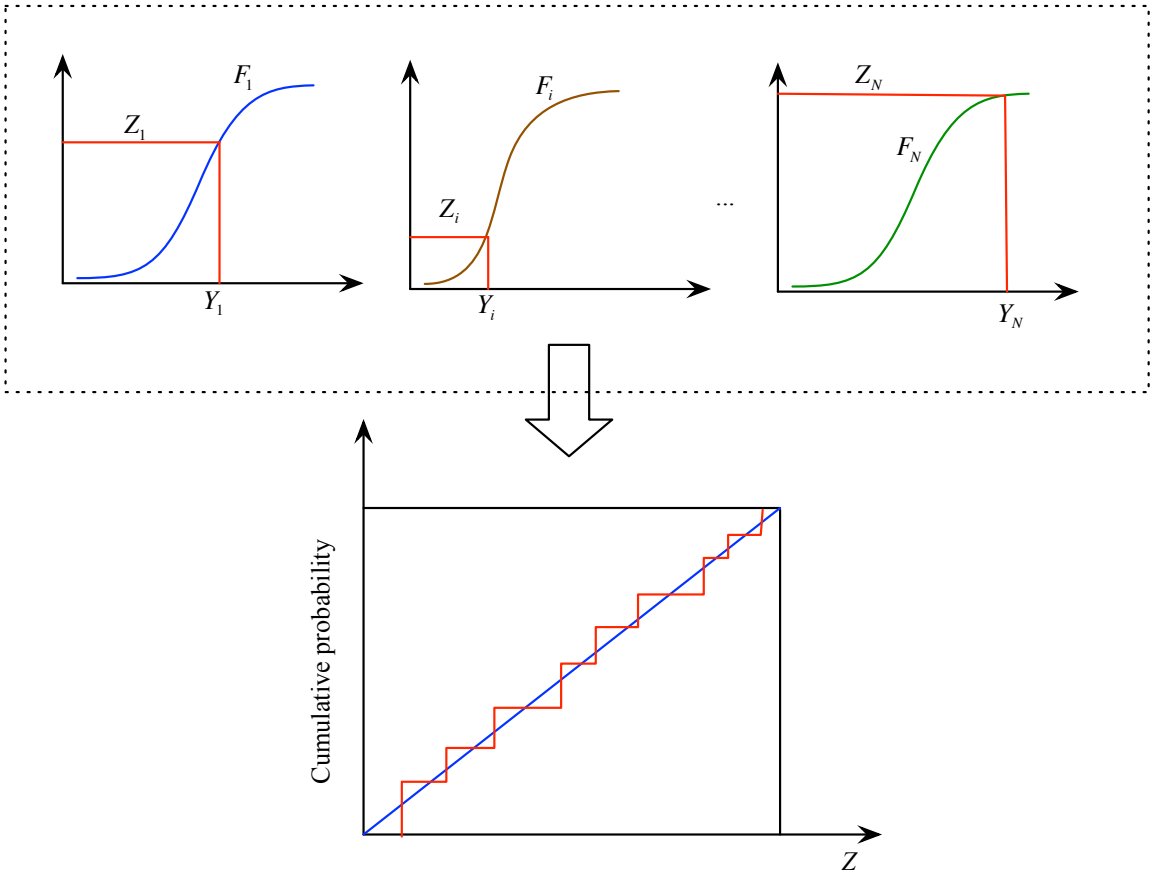


Figure 4.5 Empirical CDF of PIT from different prediction conditions

$Y_i \sim F_i$ for any predicted probability distribution, Z conforms to a standard uniform distribution, i.e. $Z \sim U(0,1)$. The empirical CDF is then compared with the ideal uniform CDF, which is a straight 45° line, and a statistical test can be employed to check the uniformity of the empirical CDF. Intuitively, if the model or modeling procedure used to provide these probabilistic predications coincides with the true data generating process, the observed realizations should be no different than randomly drawing samples from each distribution.

4.4.3 Continuous ranked probability score

Although PIT is intuitively attractive and simple to apply, it does not provide a quantitative measure using which competing models or modeling procedures can be compared. In other words, PIT is more suitable to visually check whether the predicted CDFs are compatible with observations. Continuously ranked probability score (CRPS) is a function developed from the information theory (Gneiting and Raftery 2007). Gneiting and Raftery (2007) introduce two criteria used to evaluate probabilistic predictions. They argue that

> "…the goal of probabilistic forecasting is to maximize the sharpness of the predictive distributions subject to calibration. Calibration refers to the statistical consistency between the distributional forecasts and the observations, and is a joint property of the forecasts and the events or values that materialize. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only."

CRPS provides a summary measure that addresses both sharpness and calibration simultaneously. Let *F* denote the prediction CDF and *G* denote the empirical CDF constructed from realizations. CPRS is defined as:

$$CRPS(F,G) = \int_{-\infty}^{\infty} (F(y) - G(y))^2 \, dy = \int_{-\infty}^{\infty} d^2(y) dy. \tag{4.4}$$

Figure 4.6 depicts the notes and the CPRS concept for multiple observations and for a single observation case as well. The integral of (4.4) can be simply evaluated if F is obtained from samples of Monte Carlo simulation. For single observation case, the CRPS can be simplified as:

$$CRPS(F,y_0) = \int_{-\infty}^{\infty} (F(y) - I\{y \geq y_0\})^2 dy = E_F \mid Y - y_0 \mid - \frac{1}{2} E_F \mid Y - Y' \mid, \tag{4.5}$$

where y_o is the single observation, $I\{y \geq y_0\}$ is a delta function, which takes one if $y \geq y_0$ and zero otherwise. E_F is the expectation over the distribution of F, Y' is an independent random variable with identical distribution as Y. If Y is represented by samples, then Y' can be obtained by random permutations of the sample set. From equation (4.5), it is observed that the CRPS generalizes the absolute error (AE) to which it reduces if F is a deterministic prediction, i.e. a point prediction. Thus, CRPS enables the comparison of probability predictions with point prediction.

CRPS is a very popular scoring function used to evaluate probabilistic prediction in weather forecasting (Casati et al. 2008; Sloughter, Gneiting, and Raftery 2010; Gneiting and Katzfuss 2014). It is negatively oriented and strictly proper, i.e. $CRPS(F,G) \geq CRPS(G,G)$ with equality if and only if $F = G$. Other desirable properties and mathematical proofs are referred to Gneiting and Raftery (2007). As we have mentioned before, the results of prediction verification is a joint property of model predictions and real-world observations. The CRPS explicitly shows this idea. As depicted by Figure 4.6, the values of CRPS from one observation or multiple observations are certainly different even if F is the same. This means that a probabilistic prediction is assessed based on attainable information. The discrepancy between F and G

measured by CRPS thus relates to two components: (1) the discrepancy between F and the ideal probabilistic prediction, and (2) the finite number of verification data points.

4.5 Application to simulated examples

Before applying our probabilistic prediction verification methods to real building data, it is useful to examine their efficacy on simulated data, for which we can control the true data-generating process described as follows:

Step 1: We consider 100 buildings as our sample size. The actual BEUI for each building is a random variable denoted by $B_i, i = 1,....,100$. B_i is modeled as normal distribution $B_i \sim G_i$, where $G_i = N(\mu_i, \sigma_i^2), i = 1,...,100$. The mean and variance are independently sampled from uniform distributions, $\mu \sim U(100,200)$ and $\sigma \sim U(10,20)$.

Step 2: We draw a random sample from each G_i, representing a single real-valued observation from each building. In total, we have 100 observations.

We will examine the usefulness of the PIT and CRPS methods in assessing four probabilistic predictions that differ from the data-generating process. We consider all model predictions take a Gaussian distribution as the form of their probabilistic

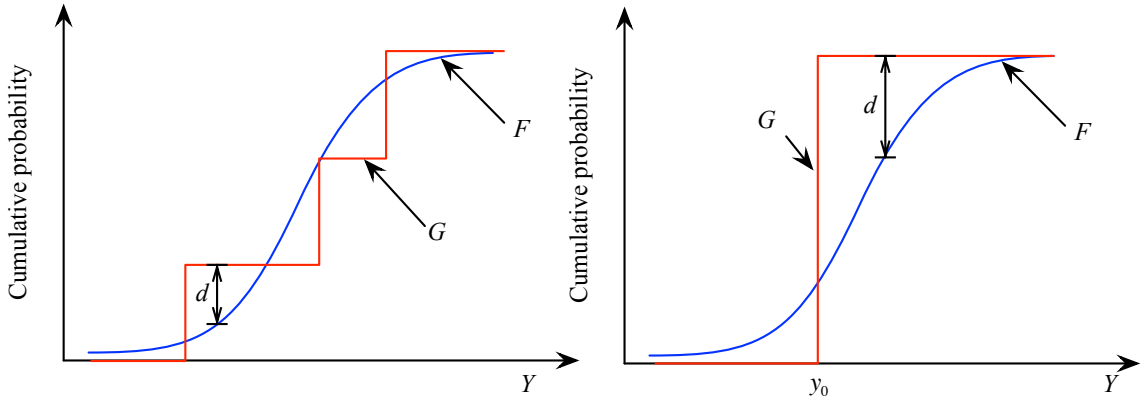


Figure 4.6 Continuous ranked probability score for multiple observations (left) or single observation (right)

predictions, denoted by $F^j, j = 1,...,4$, yet their predicted means and standard deviations differ from the data-generating process. Consider the 100 buildings are from 100 different design specifications and are thus modeled differently, yet all models are created under one modeling procedure. Table 4.3 lists the modeling procedures and the probability distributions. Specifically, MP-0 is the ideal modeling procedure that provides all 100 distributions coinciding with the data-generating process; modeling procedure 1 (MP-1) offers biased predictions due to underestimating the means; MP-2 offers biased predictions due to overestimating the means; MP-3 provides unbiased predictions, yet underestimates the standard deviations; MP-4 also provides unbiased predictions, yet overestimates the standard deviations.

4.5.1 PIT results: one observation for each building model

Figure 4.7 to Figure 4.11 plot the empirical CDFs from 100 samples of PIT for MP-0 to MP-4. As we have expected, Figure 4.7 shows that the empirical CDF of PIT from MP-0 approximates the 45° line, which is the ideal uniform CDF. The empirical CDFs of PIT from MP-1 and MP-2 substantially deviate from the uniform CDF as shown by Figure 4.8 and Figure 4.9. Specifically, because MP-1 provides biased predictions due to underestimates in the means, the shape of the empirical CDF of PIT shown in Figure 4.8 indicates that most PITs evaluated at the observation point are higher than 0.5, i.e.,

Table 4.3 Design of simulated examples

Model Index	**Prediction**	**Note**
MP-0	$G_i = N(\mu_i, \sigma_i^2)$	Data-generating process
MP-1	$F_i^1 = N(0.9\mu_i, \sigma_i^2)$	Biased due to underestimate
MP-2	$F_i^2 = N(1.1\mu_i, \sigma_i^2)$	Biased due to overestimate
MP-3	$F_i^2 = N(\mu_i, (0.3\sigma)^2)$	Underdispersed
MP-4	$F_i^2 = N(\mu_i, (3\sigma)^2)$	Overdispersed

the medians of the predictive distributions. The mean of the PIT from MP-1 is also depicted in Figure 4.8, which is about 0.9. Contrary to MP-1, MP-2 provides biased predictions due to overestimate. The shape of the empirical CDF of PIT in Figure 4.9 shows most PITs are lower than 0.5. This means that the observation is more likely to appear towards to the left half of the predicted distribution.

The results from MP-3 and MP-4 show empirical CDF of PIT when the variance or uncertainty cannot be accurately predicted. Obviously, both empirical CDFs substantially deviate from the perfect uniform CDF, i.e. the 45° line. Furthermore, Figure 4.10 shows that there are fewer PITs in the middle than being pushed towards the two ends. This corresponds to the fact that the uncertainty is underestimated by MP-3 such that the narrow range of predictions may not cover the observation point. In contrast, MP-4 provides an overdispersed prediction, so the observation point is not likely to appear near to the bounds of the prediction. Thus, many observations points appear in the middle range as shown in Figure 4.11. In summary, the PIT method not only provides

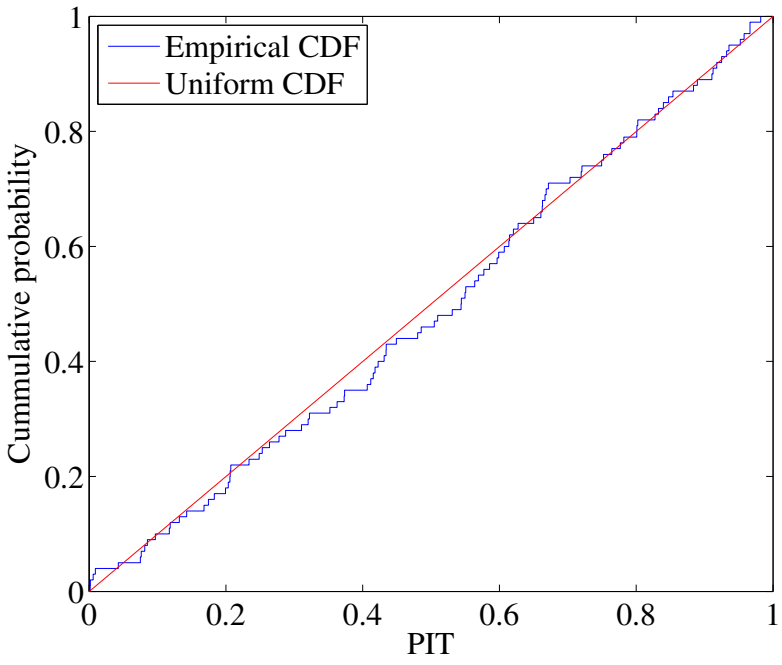


Figure 4.7 Empirical CDF of MP-0

probabilistic verification, but also has certain diagnostic capabilities.

4.5.2 PIT results: multiple unequal observations for each building model

In building practice, we may obtain multiple observations corresponding to a probabilistic prediction if we have identical buildings coming from the same design specification. More generally, different probabilistic predictions may have a different number of observation points used for verification. This section aims to test whether the preceding results hold under multiple unequal observations. In our simulated test, instead of sampling only one from the data-generating process as our virtual observation, we sample m times. m *is* a random integer between 1 to 10. In total, we will obtain a total 514 observations with different numbers of observations from different probabilistic predictions. Figure 4.12 to Figure 4.16 show the results of empirical PIT with total sample size $N = 514$. The results derived from multiple unequal observations provide the same findings compared with the one observation per building model example. Because of larger sample size, the empirical CDFs are smoother than the previous case.

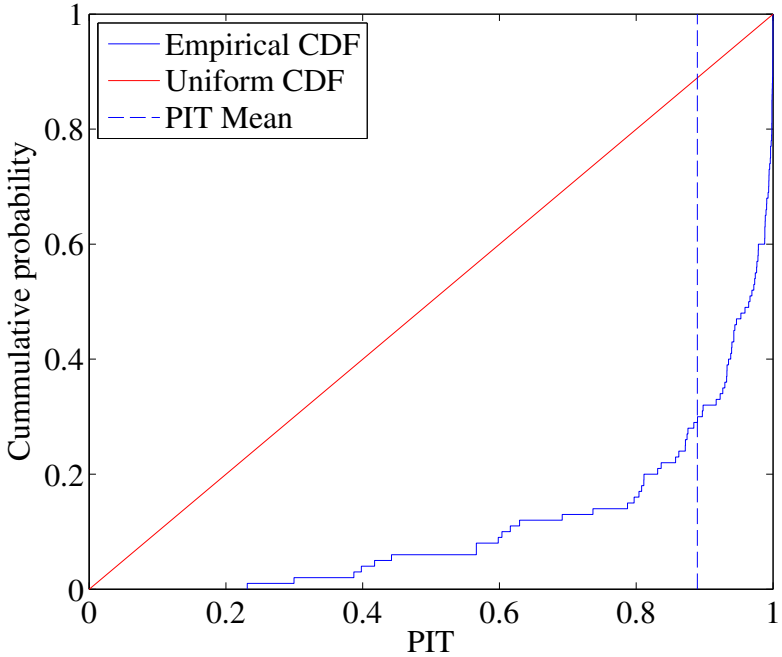


Figure 4.8 Empirical CDF of MP-1

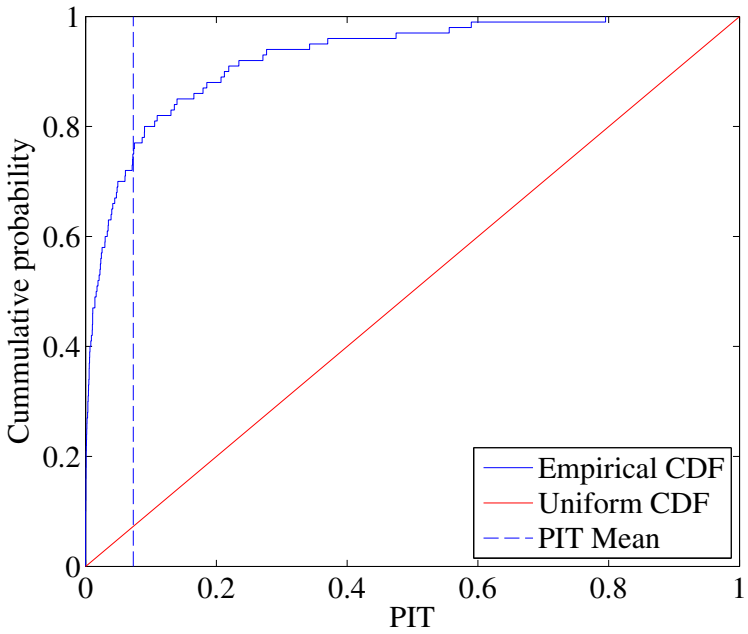


Figure 4.9 Empirical CDF of MP-2

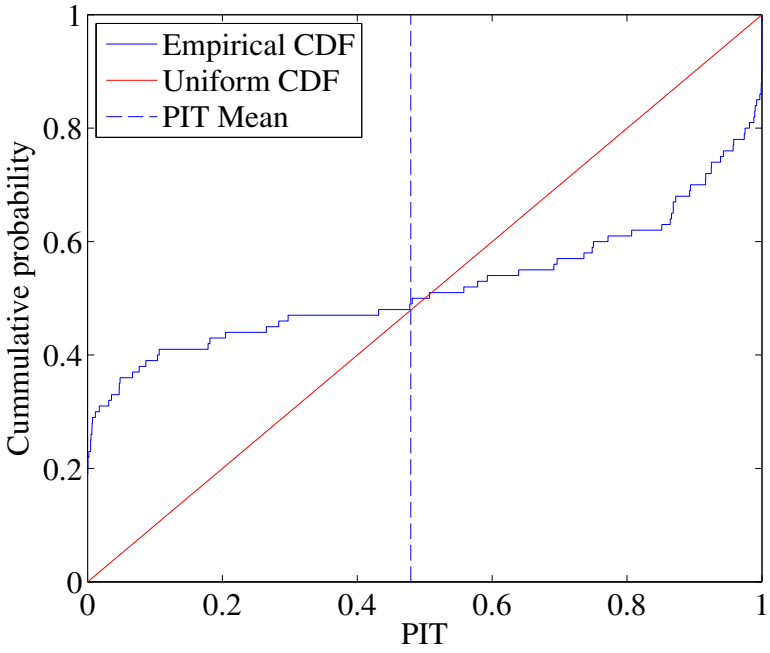


Figure 4.10 Empirical CDF of MP-3

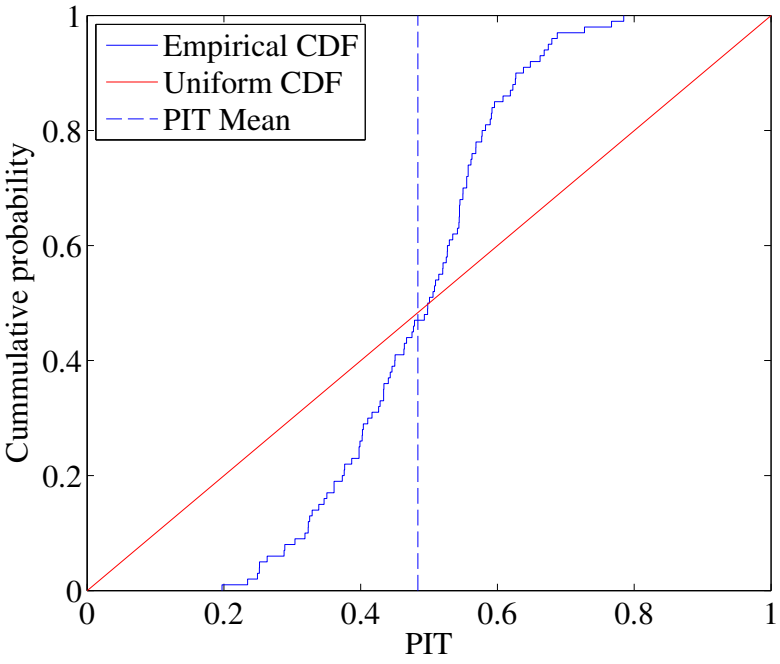


Figure 4.11 Empirical CDF of MP-4

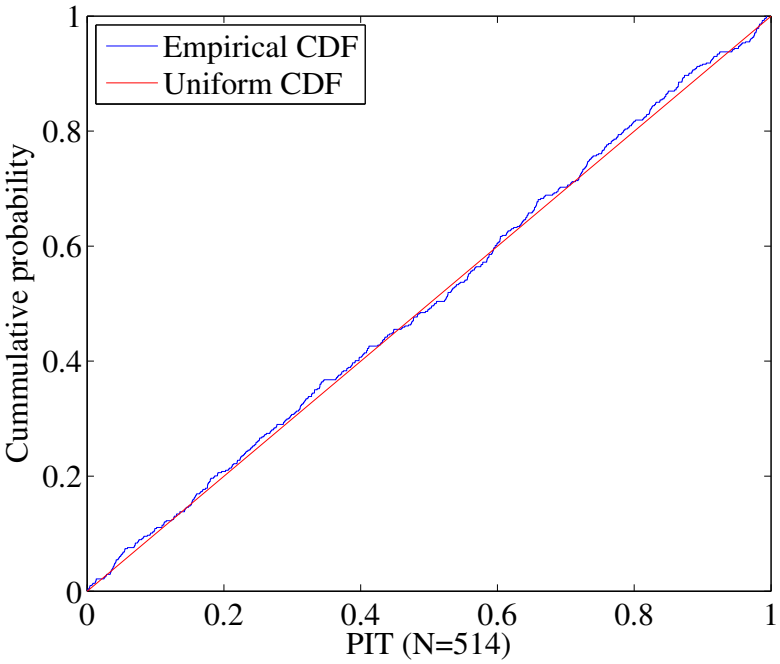


Figure 4.12 Empirical CDF of MP-0: multiple unequal observations

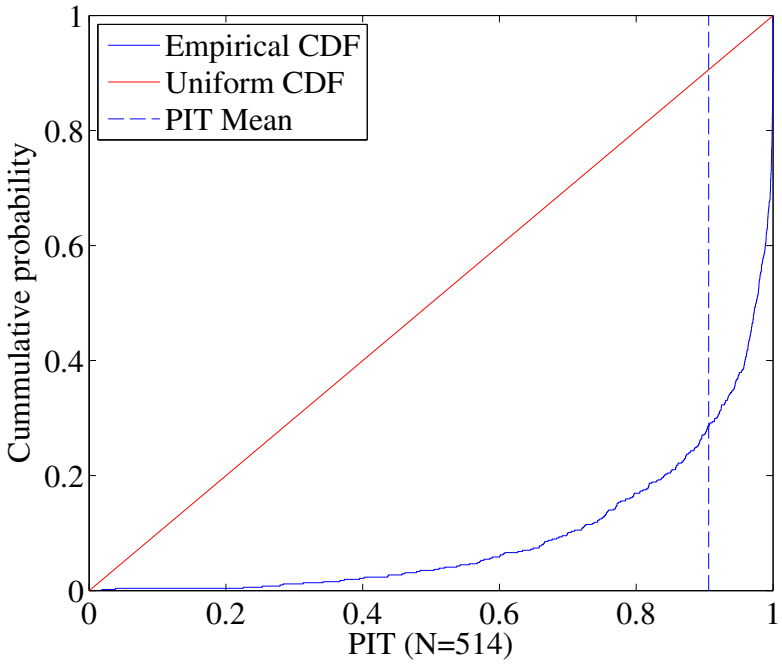


Figure 4.13 Empirical CDF of MP-1: multiple unequal observations

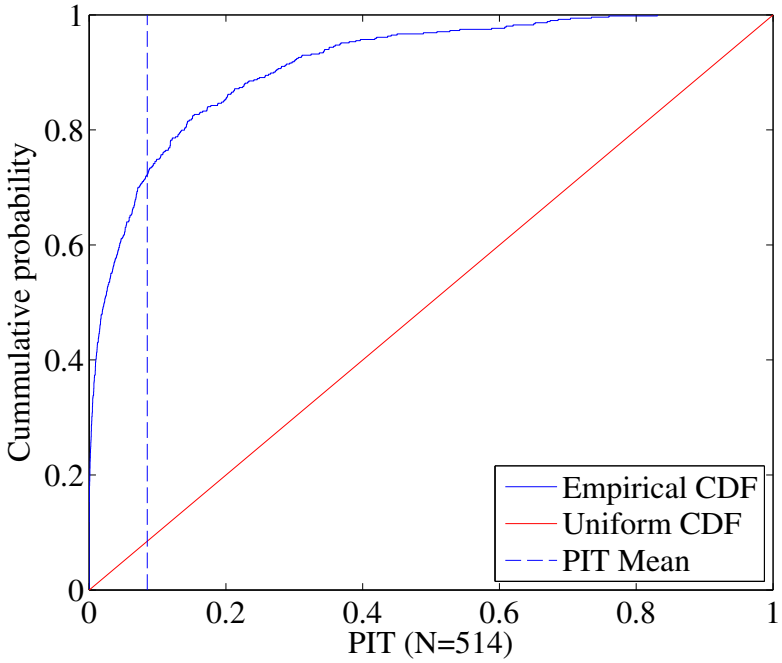


Figure 4.14 Empirical CDF of MP-2: multiple unequal observations

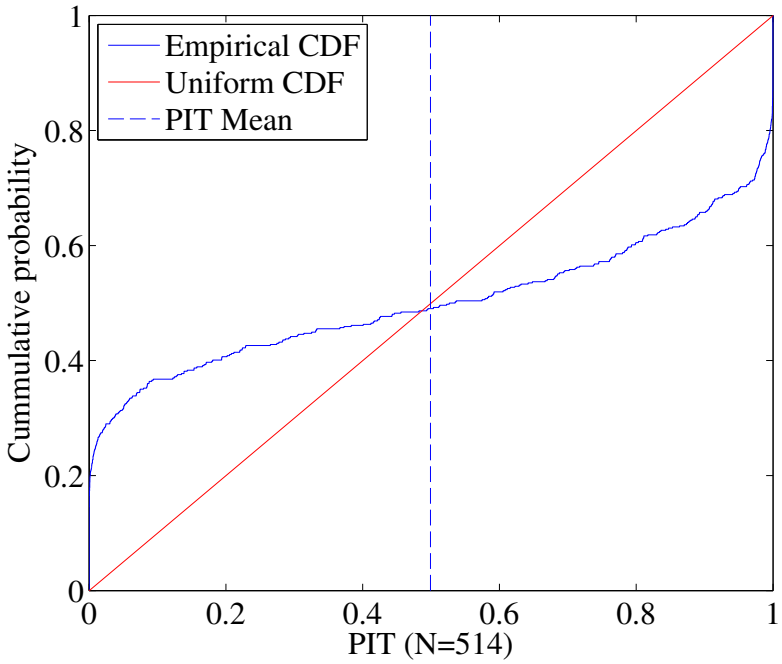


Figure 4.15 Empirical CDF of MP-3: multiple unequal observations

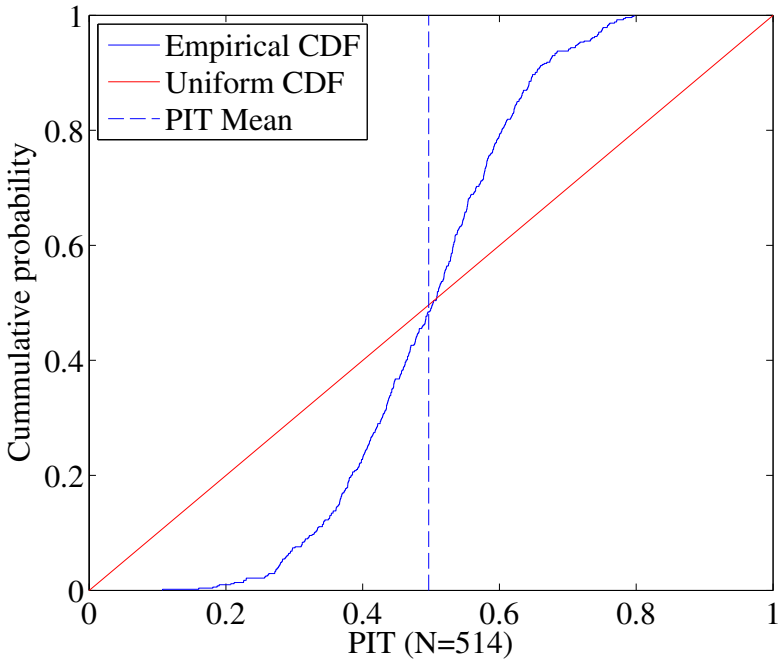


Figure 4.16 Empirical CDF of MP-4: multiple unequal observations

4.5.3 CRPS results

Continuous ranked probability score provides a summary measure for the model prediction performance, which is suitable to rank competing models. First, we show the results of the CRPS score for 100 building models individually. We start from using one single observation for each building model. We compare the CRPS scores calculated by MP-0 (i.e. the ideal model) and by MP-1, and MP-2 in Figure 4.17. The straight line $x = y$ is plotted as the reference. Points falling above the line are the cases where the ideal model scored better than the other. Points falling below the line are the cases that the non-ideal model scored better. We have in total 11 (4 from MP-1 and 7 from MP-2) out of 200 of such points when the non-ideal model happens to obtain a better score than the ideal one. As explained previously, model prediction verification is the joint property of the model under study and the events or values that materialize. Thus, it is possible that the non-ideal model obtains a better score than the ideal one by chance. Figure 4.18 depicts an example when the single observation point happens to be closer to the mean of MP-3 than MP-0, so MP-3 obtains a better CRPS score than MP-0. Following this idea, if

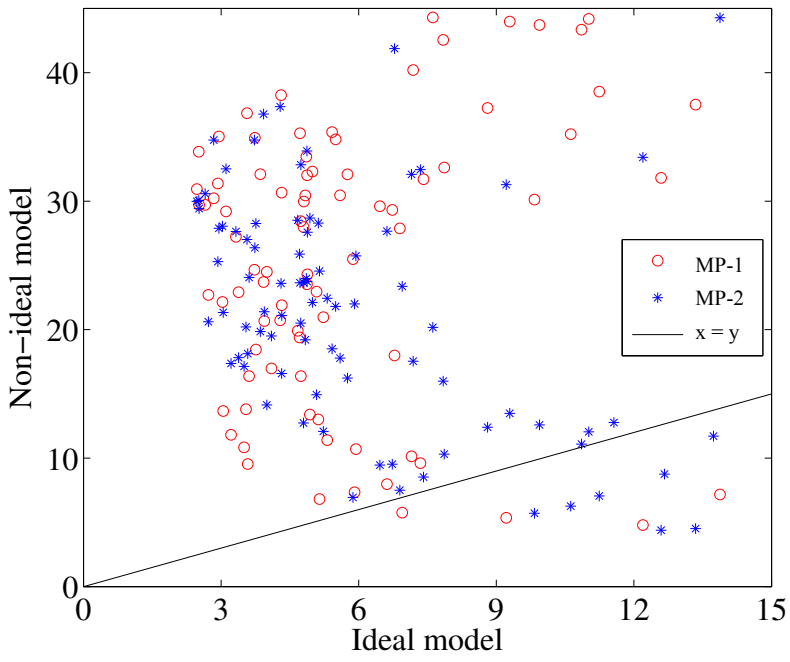


Figure 4.17 CRPS MP-0 vs. MP-1 and MP-2; single observation for each building

there are more observation points corresponding to each probabilistic prediction, the chance of scoring non-ideal models better than the ideal-model decreases. Figure 4.19 shows the comparisons using 20 observation points per building model. Figure 4.20 and Figure 4.21 plot the comparisons between MP-0 and MP-3, and MP-4 for 1 and 20 observation points. The general findings are similar to the comparisons with MP-1 and MP-2.

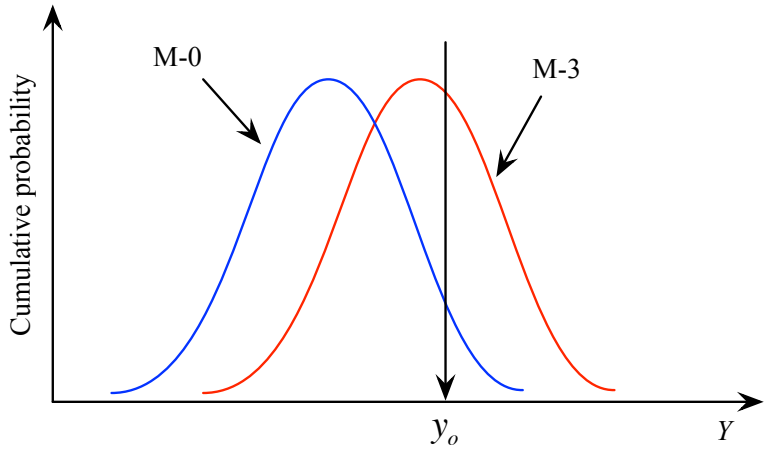


Figure 4.18 MP-3 obtains better score than MP-0 when y_o is observed

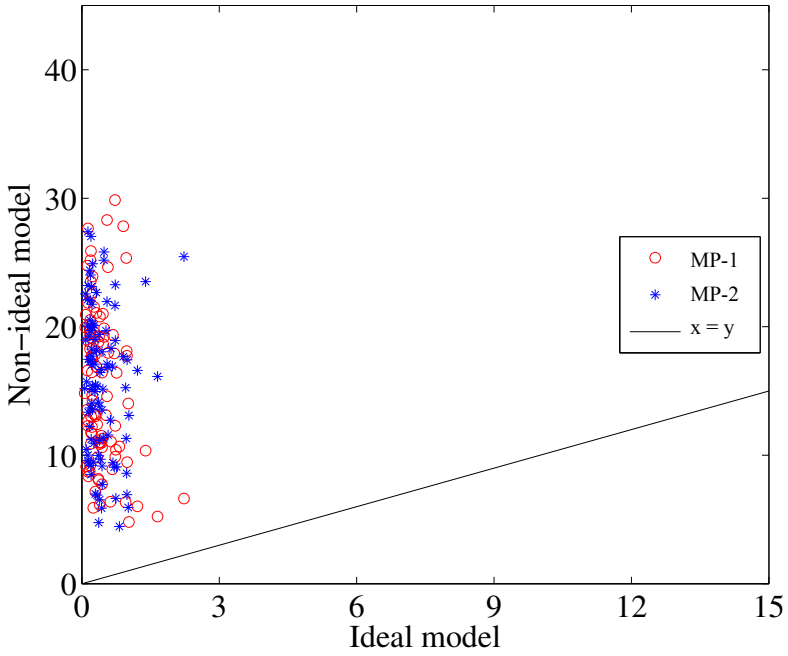


Figure 4.19 CRPS MP-0 vs. MP-1 and MP-2; 20 observations for each building

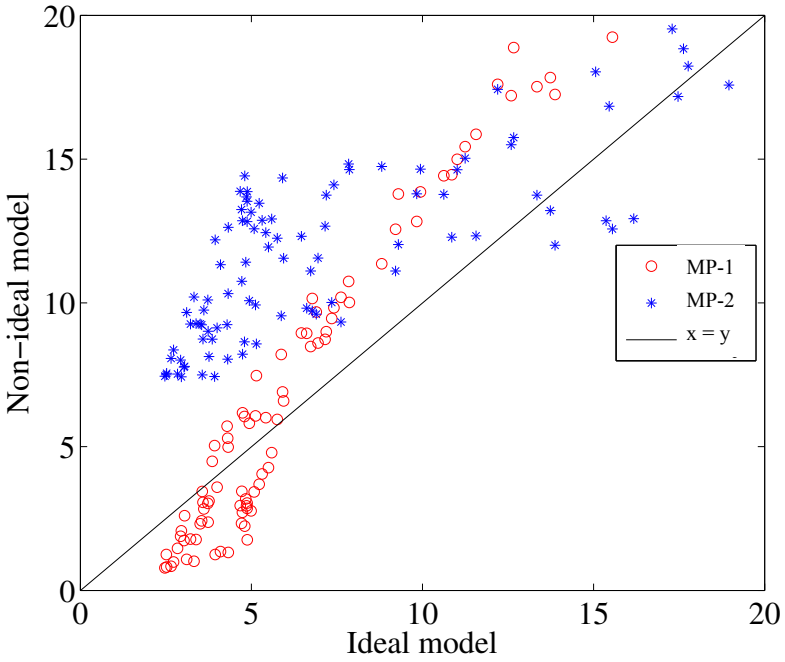


Figure 4.20 CRPS MP-0 vs. MP-3 and MP-4; single observation for each building

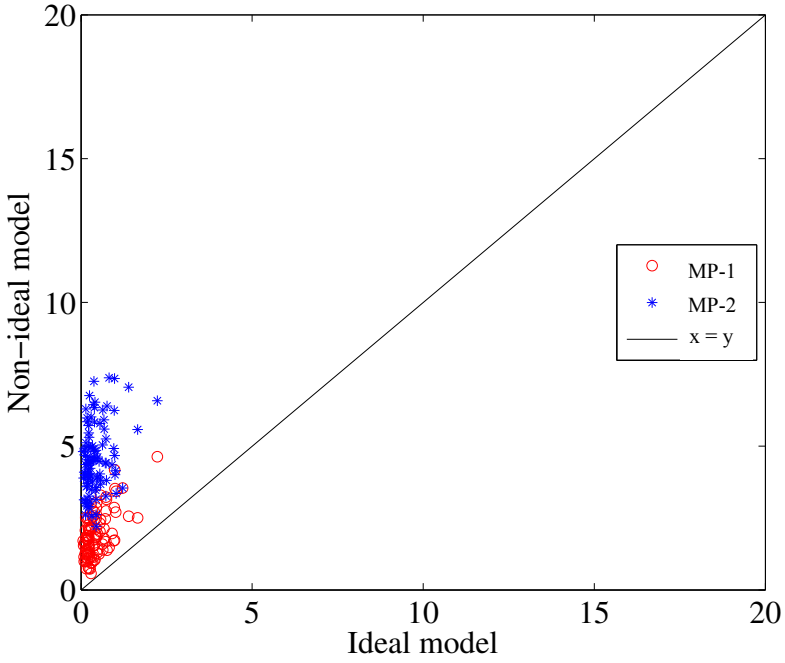


Figure 4.21 CRPS MP-0 vs. MP-3 and MP-4; 20 observations for each building

Table 4.4 summarizes the mean CRPS and the rankings of the model. As the number of observation points increase, the CRPS for the models decreases. The CRPS for

the ideal model will eventually approach to zero as we use more than more data points. As for the rankings of each model, as we have expected MP-0 ranks the top even at one-point case. It means that either using more building models or more observations at each model will help to rank the models. Because MP-1 and MP-2 have similar performance evaluated by CRPS, their relative rankings may fluctuate from one case to another. In practice, if models or modeling procedures have similar prediction performance, one is then free to decide whichever one prefers.

These examples show that the CRPS can discern a non-ideal model from the ideal one due to either the error in means or in variance. To illustrate how the CRPS simultaneously consider the two types of errors, we conduct a follow-up simulation study. In this study, we consider a probabilistic prediction from model as $F_i = N(\alpha\mu_i^T, \gamma\sigma_i^T)$, whose mean and standard deviation deviate from those of the data generating process by factors of α and γ. Figure 4.22 plots the distribution of the

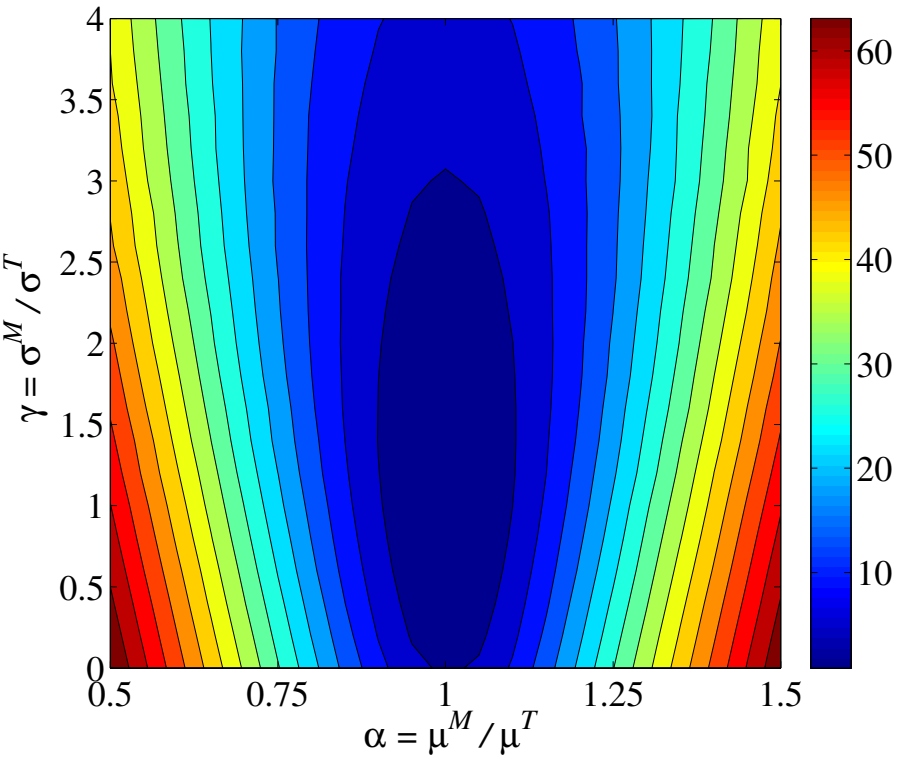


Figure 4.22 Expected continuous ranked probability score

expected CRPS over different values of α and γ. Obviously, the minimum CRPS is realized when both α and γ are close to 1. In other words, the CRPS can simultaneously penalize a probabilistic prediction with errors either in mean or variance.

Table 4.4 Mean continuous ranked probability score for the simulated examples

Model Index	**One Point**		**20 Points**	
	CRPS	Ranking	CRPS	Ranking
MP-0	8.1	1	0.4	1
MP-1	24.6	5	15.4	4
MP-2	23.6	4	15.7	5
MP-3	9.4	2	1.9	2
MP-4	12.0	3	4.5	3

CHAPTER 5 PROPABILISTIC PREDICTION FOR CLOSING THE BUILDING ENERGY PERFORMANCE GAP: CASE STUDIES

5.1 Case buildings

Six buildings on the campus of the Georgia Institute of Technology are used to evaluate the methodology for closing the building energy performance gap. The general information of six buildings is listed in Table 5.1. All six buildings are mixed use, primarily consisting of offices, classrooms, and labs. All buildings have mechanical cooling and heating systems for air-conditioning and ventilation. The cooling and heating sources come from chilled water and steam supplied by campus central plants through district cooling and heating systems.

Table 5.1 Building general information summary

Name	**Space function**	**Built/major renovation year**	**Gross floor area (m^2)**
Manufacturing Related Disciplines Complex (**MRDC**)	Office, classroom, lab, light industry, computer lab	1995	11,000

Cherry L. Emerson Building (**Cherry**)	Office, research lab	1959	5,500
Centennial research building (**CRB**)	Office, conference room, research lab	1984	17,700
College of Computing (**COC**)	Office, conference room, research lab	1989	10,700
Hinman Research Building (**Hinman**)	Office, design studio/classroom, computer lab	1939/2010	3,200
Student Health Center (**Health Center**)	Office, medical room	2002	3,500

Figures 5.1 to 5.6 show the elevations and typical floor plans for six buildings. The first case building is the Manufacturing Related Disciplines Complex (MRDC). The building houses faculty offices, classrooms, computer labs, and light industrial equipment for research experiments. It was built in 1995 with a gross floor area of 11, 000m^2, it has not gone though major renovation since then. It has fluorescent lighting and nominal office and lab plug-in appliances. The Cherry L. Emerson Building was the second case building in this study. It is a three-story building, built in 1959, and has a gross floor area of 5, 500m^2. The core zones are typically used for chemistry and biology labs, and the perimeter zones are used as faculty or graduate student offices. The Centennial research building (CRB) is a 6-story, 17,700m^2 facility with space for electronic research and other high-security research activities. It was built in 1984. The College of Computing Building (COC), a 10,700m^2 facility built in 1989, houses administrative offices, classrooms, research labs, meeting space for undergraduate and graduate student organizations. The Hinman Research Building (Hinman), designed in 1939, is one of the historical buildings on campus. It was recently restored by the College of Architecture in 2011. The restoration preserved the original materials and important interior and exterior architectural features. Currently, the building houses about 3,500m^2 of multi-use space for graduate level architecture studios (mostly in the high-bay area), computer labs, research labs, administration offices, and galleries. The original concrete and steel construction was retrofitted to a LEED Gold standard of sustainability. The Student Health Center (Health Center), built in 2002, is a relatively new constructed building

among our case buildings. It provides 3,500m^2 space for health care services including offices and medical rooms.

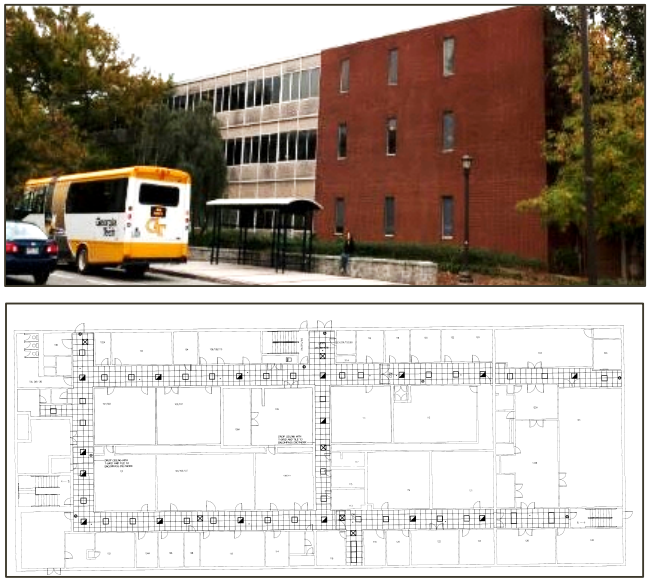


Figure 5.2 Elevation and typical floor plan of the second case building: Cherry L. Emerson Building

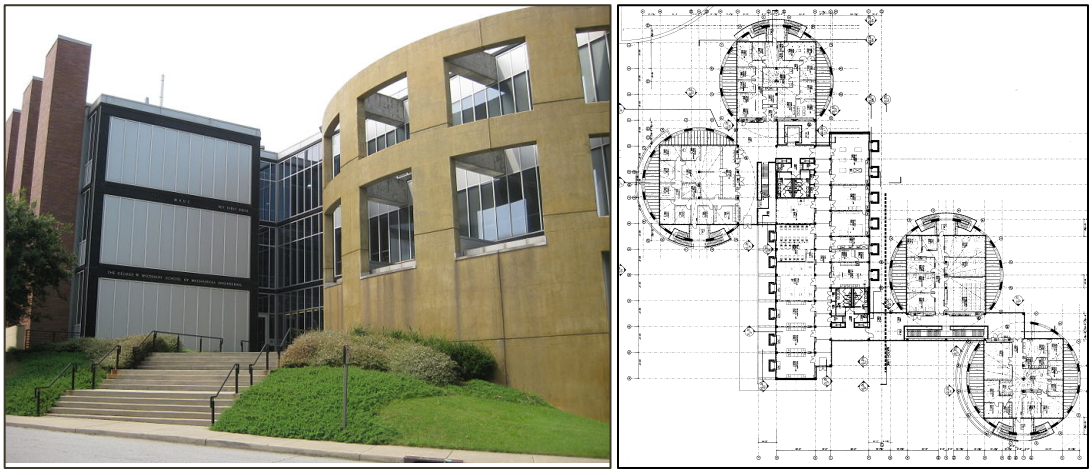


Figure 5.1 Elevation and typical floor plan of the first case building: MRDC

Figure 5.3 Elevation and typical floor plan of the third case building: CRB

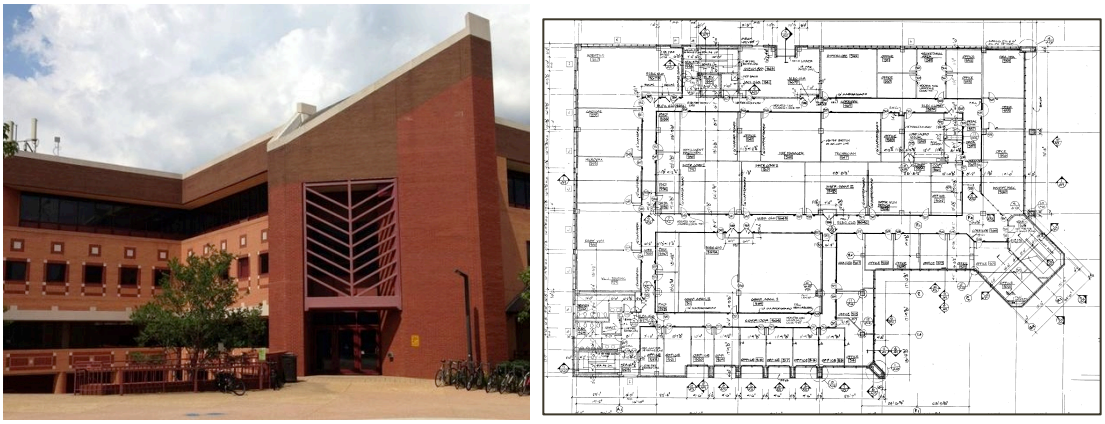


Figure 5.4 Elevation and typical floor plan of the fourth case building: COC

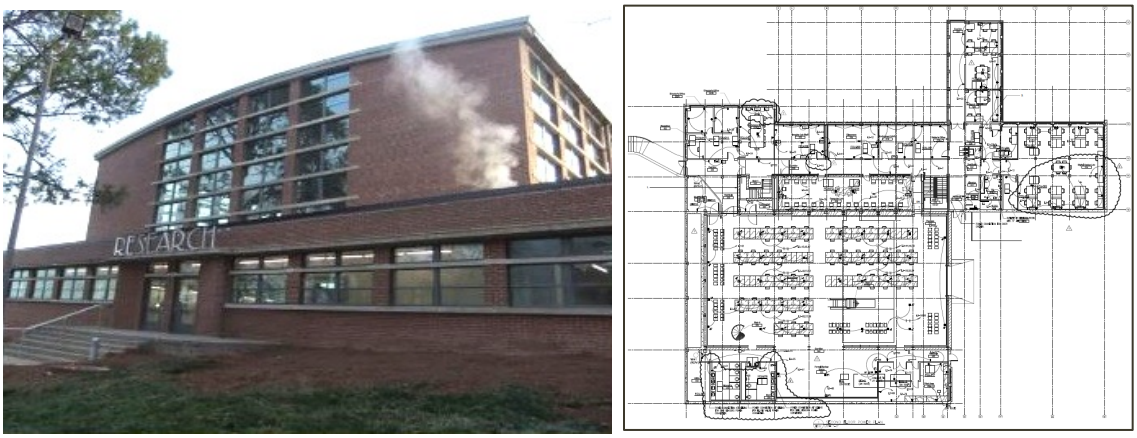


Figure 5.5 Elevation of typical floor plan of the fifth case building: Hinman

Figure 5.6 Elevation of typical floor plan of the sixth case building: Health Center

5.2 Energy consumption measurement

Energy consumption is classified into three categories: district cooling thermal energy, district heating thermal energy, and electricity. However, in the current measurement system, only district cooling and electricity are well metered. Moreover, since the building electricity consumption excludes cooling plant systems (e.g., chiller and cooling tower), the electricity consumption data dominated by lighting loads, plug loads and process loads are not suitable to test the accuracy of the dynamic simulations whose primary use pertains to the prediction of HVAC systems electricity consumption. Thus, this study focuses on the district cooling energy use for prediction verification purpose.

Figure 5.7 shows a typical diagram of chilled water control and measurement system for campus buildings. The chilled water energy is metered by three measured variables: inlet chilled water temperature, outlet chilled water temperature, and the water flow rate. Three measured variables and the calculated chilled water thermal energy are recorded using a data logger at a 5-minute interval. The high-frequent measurement data are then aggregated into monthly and total annual consumption. The control and measurement systems were installed and commissioned by Schneider Electric in 2011 and 2012. Because both temperature and flow rate are relatively easy to measure with normal measurement technology, systematic (biased) measurement errors should be acceptable compared to the level of accuracy needed to examine the modeling results if the installation and commissioning processes are performed adequately, and the sensors are adequately calibrated. Because we primarily focus on monthly and annual energy use, the random measurement errors become less an issue after aggregating hundreds or thousands of measured points. More discussion about data quality is presented in the final section of this chapter.

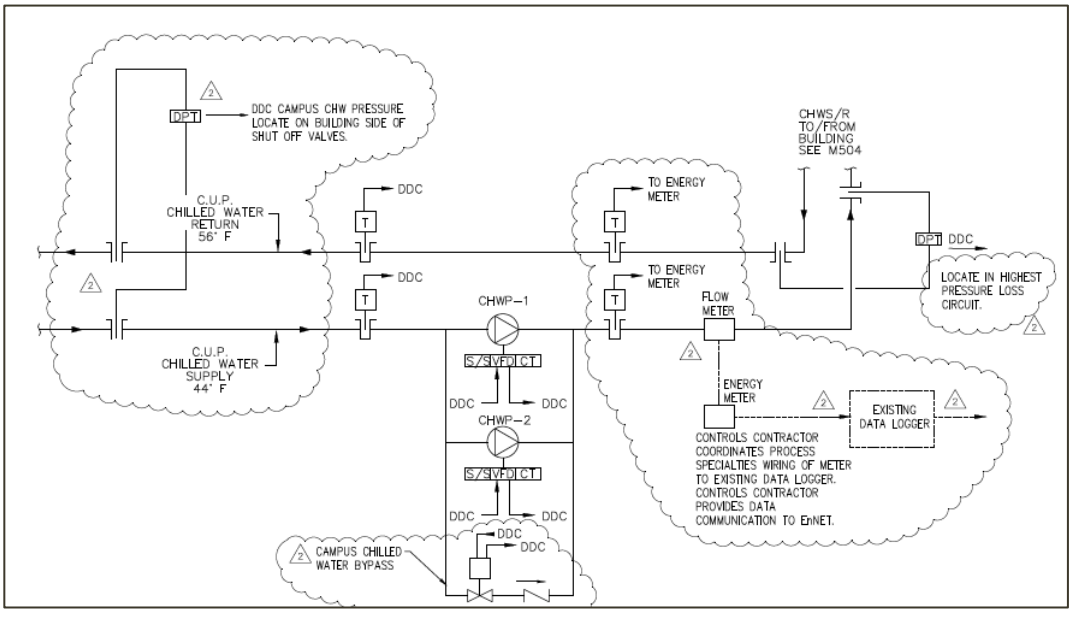


Figure 5.7 Chilled water control and measurement system

5.3 Building energy models

We follow the standard modeling guidelines of the ASHRAE 90.1-2007 (ASHRAE 2007b) in modeling the buildings in EnergyPlus V7.0. The main modeling strategies and model parameters can be broadly summarized within the following groups: (1) weather data, (2) building geometry, (3) thermal zoning, (4) construction and material properties, (5) HVAC systems, (6) internal loads (lighting, plug-in appliances, occupants), and (7) building operation. Table 5.2 summaries the primary modeling strategies and model parameters. Regarding lighting and plug-load, two options are explored in our study. The following section introduces the purpose.

5.4 Experimental design of modeling procedures

Six buildings are modeled by five modeling procedures as follows. Modeling procedure 1 (MP1) closely follows the guidelines of the standards (ASHRAE 2007a). The standardized lighting and plug load density and schedules comes from the ASHRAE standards the DoE reference building models, the medium-size office in particular (Deru et al. 2011). MP1 runs a single deterministic simulation, offering the best (point) prediction of building energy consumption. The performance of MP1 forms as the baseline to evaluate the other modeling procedures.

Modeling procedure 2 (MP2) uses the Georgia Tech Uncertainty and Risk Analysis Workbench (GURA-W) and the UQ repository to quantify the model input parameter uncertainty and to execute the Monte Carlo simulations. We further divide MP2 by MP2a and MP2b. MP2a runs the Monte Carlo simulations using the probability distributions of input parameters, whereas MP2b only runs a single deterministic simulation using the means of the probability distributions of input parameters as the best estimates of these uncertain parameters. Similar to MP1, MP2b also provides a point prediction for any given month or year. The differences are only in the values of lighting, plug load density, and the hourly schedules. The difference in simulation results between

MP1 and MP2b is then attributable to the effect the differences in lighting and plug loads introduced in Chapter 3.

Modeling procedure 3 (MP3) uses the full capability of the GURA-W and the UQ repository in that MP3 also considers the model form uncertainty. Since the model form uncertainty modified some simulation algorithms in the current EnergyPlus V7.0 and thus implemented through the programming of the EnergyPlus source codes in Fortran, we denote the modified simulation platform as the Georgia Tech version of EnergyPlus, i.e., GTPlus. GTPlus is the simulation engine, using the full functionality of the Georgia Tech Uncertainty and Risk Analysis Workbench (GURA-W). We further divide MP3 into MP3a and MP3b. The former runs Monte Carlo simulations, whereas the latter runs a single deterministic simulation using the means of uncertainty parameters.

Table 5.2 Building modeling and uncertainty quantification

Category	**Model parameters/ modeling strategy**
1. Weather	2012 and 2013 AMY
2. Building geometry	Design specifications
3. Thermal zoning	Detailed zoning based on HVAC mechanical design specifications
4. Construction and material properties	Design specifications
5. HVAC systems	HVAC mechanical design specifications
6. Internal loads	Building averaged peak use; Building averaged weekday, weekend, and holiday hourly schedules
6.1. Lighting and plug load peak use (W/m2)	1. ASHRAE Standards 2. A measurement dataset from ASHRAE research project 1093-RP 2001
6.2. Lighting and plug load hourly schedules	1. Office standardized schedules from ASHRAE standards 2. A measurement dataset from the ASHRAE research project 1093-RP 2001
7. Building operation	Standard hourly occupant density and schedules

In summary, we design five modeling procedures. MP2a and MP3a run Monte Carlo simulations; MP1, MP2b and MP3b run deterministic simulations. Correspondingly, two types of probabilistic predictions on the building cooling energy use are obtained from MP2a and MP3a. Meanwhile, probabilistic predictions can also be generalized to point prediction cases by using the mean or median of the Monte Carlo simulations. Therefore, we will obtain five different point predictions of building cooling energy use from the five modeling procedures.

Simulation results from each modeling procedure are to be compared with measurements to calculate the discrepancies using statistical measures introduced in Chapter 4. Moreover, the comparisons of the discrepancies among different modeling procedures will show the effect on improving the model predictions. For example, comparing MP1 against MP2b shows the effect of lighting and plug loads; comparing MP2a against MP3a shows the effect of model form uncertainty quantification on the probabilistic predictions of district cooling energy use.

5.5 Results of the Cherry L. Emerson building

This section reports the detailed results using Cherry case building. The idea is to prepare the reader to comprehend the follow-up statistical analysis, where we derive the final conclusions using all the data from six buildings.

First, let us present the monthly results. Figure 5.8 plots the results of metered and deterministically predicted cooling energy consumption. In total, there are 24 data points corresponding to measurements from 24 successive months from January 2012 to December 2013. The simulation results plotted in Figure 5.8 include MP1, MP2b, and MP3b. Each runs a deterministic simulation using 2012 and 2013 AMY data, respectively. In general, we observe more or less differences between metered and simulated results. There are a few data points, such as 07/2012 and 08/2012, the predicted results by MP3b agree very well with metered values. In contrast, there are also cases

with visible discrepancies regarding to any model predictions, such as 11/2012 and 12/2013. As we have expected, the performance gap changes from month to month for any modeling results. Figure 5.8 shows there is not such an absolute winter that provides the best prediction in any month. In general, MP3b seems to perform better than the others. In the follow-up statistical analysis, we will offer a quantitative statistical measure, which can be regarded as a summarized evaluation over a collection of data points.

Figure 5.9 and Figure 5.10 show the probabilistic predictions of MP2a and MP2b. Both MP2a and MP2b provide 300 simulation results, each coming from one Monte Carlo simulation run. The probabilistic prediction is usually visualized with probability distribution function (PDF) or cumulative distribution function (CDF). If so, we need to display 24 plots for each individual month. To offer a compact presentation, we show the means and the uncertainty ranges for the monthly results. The uncertainty band is two times the standard deviation, which approximately covers 95% of the points in the middle range. We observe from the figures that some measurements fall in the predicted intervals, whereas some measurements, especially during winter season, fall outside these

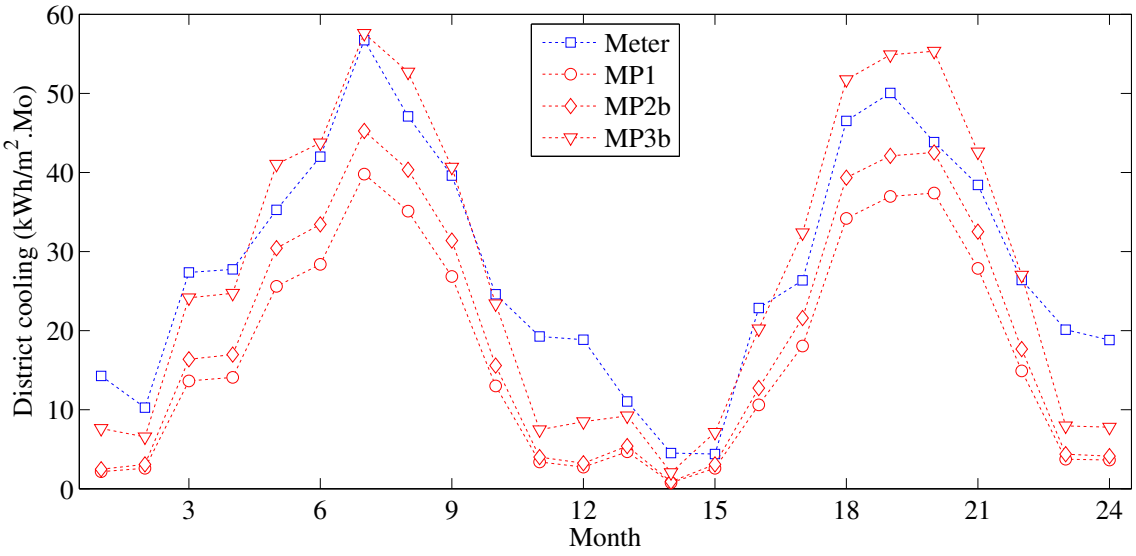


Figure 5.8 Monthly predictions of MP1, MP2b, MP3b and measurements from 01/2012 to 12/2013 for the Cherry building

intervals. For those outside points, we suspect that either the model prediction is biased or the uncertainty is underestimated. When the measurement data fall in the predicted intervals, we may gain certain confidence to the model predictions. However, one observation point is not enough to fully evaluate the probabilistic prediction. This is the reason that we shall pool those incompatible pairs of probabilistic prediction and one real-valued observation from different prediction contexts, so that we can evaluate the underling model or modeling procedure. MP2a and MP3a provide different probabilistic predictions, but it is not apparent which one performs better than the other.

The annual results are obtained by aggregating the 12 months over each year. Figure 5.11 and Figure 5.12 show the CDF of probabilistic predictions of MP2a and MP3a and the corresponding point annual measurements. Each building offers 2 model evaluation data pairs, i.e. one for 2012 and the other for 2013. Compared with MP2a, MP3a seems perform better since the metered points fall in the middle range of the probabilistic prediction of MP3a, whereas they fall towards the upper tails of MP2a.

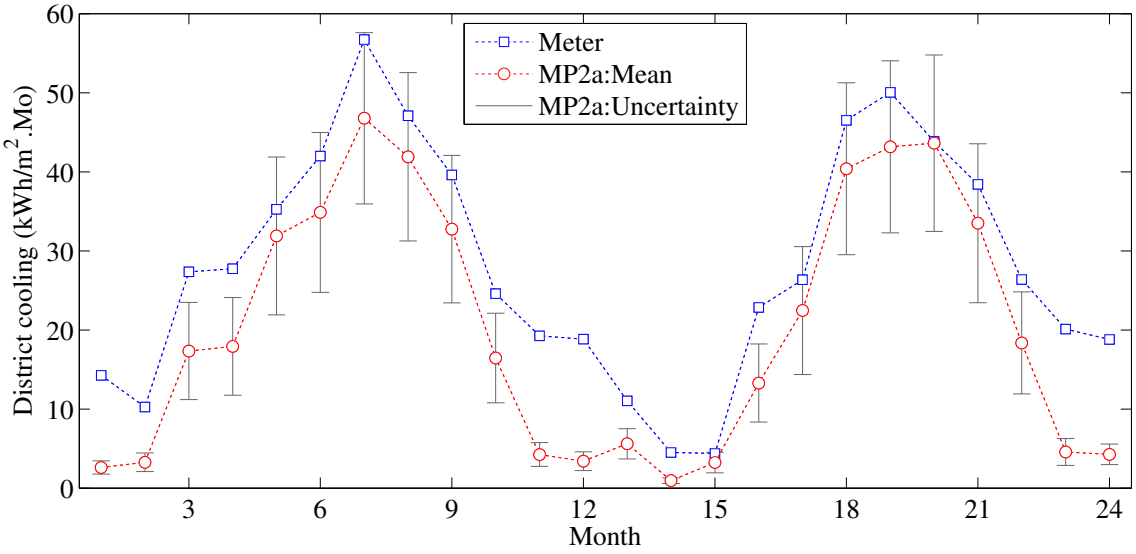


Figure 5.9 Predicted means and uncertainties by MP2a, and measurements for the Cherry building from 01/2012 to 12/2013

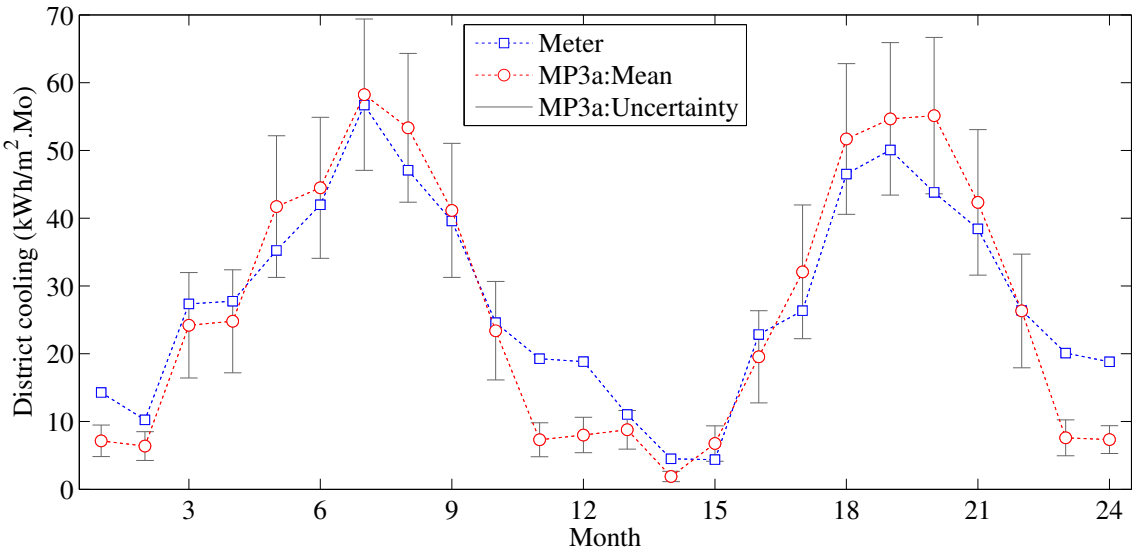


Figure 5.10 Predicted means and uncertainties by MP3a, and measurements for the Cherry building from 01/2012 to 12/2013

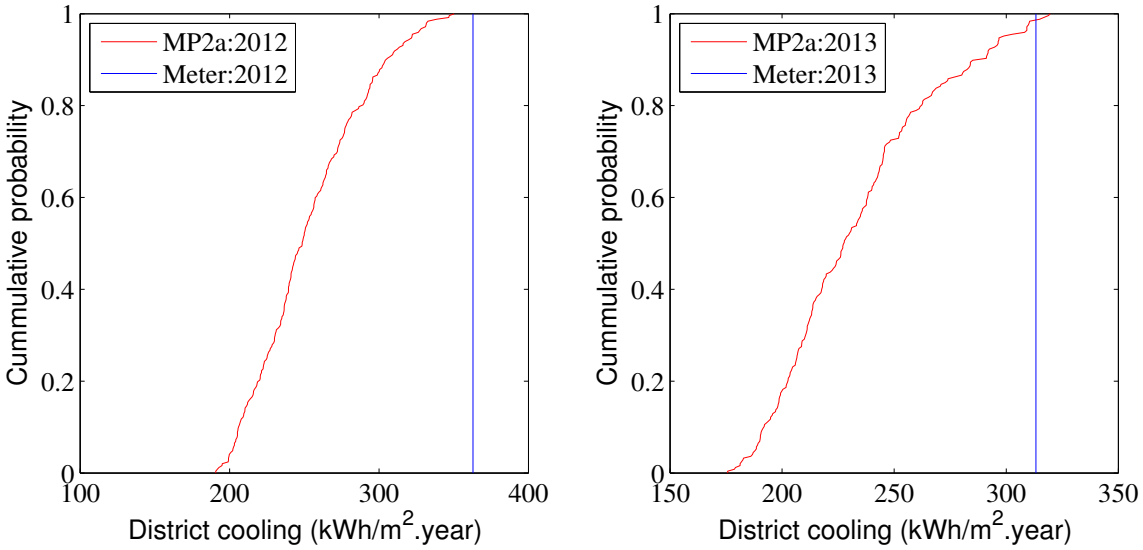


Figure 5.11 Predicted CDF by MP2a and measurements for the Cherry building: the year of 2012 (left) and the year of 2013 (right)

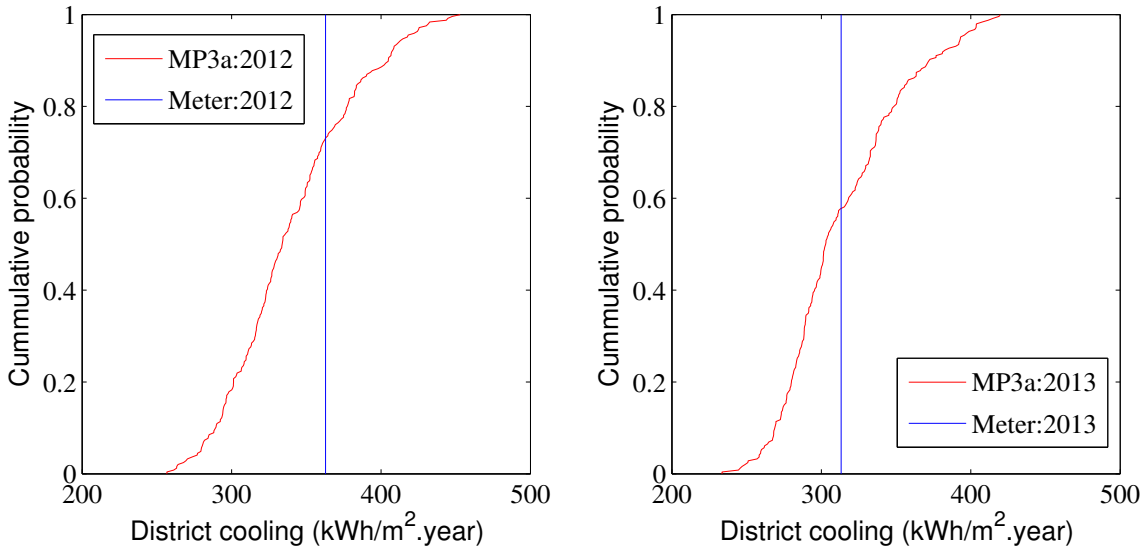


Figure 5.12 Predicted CDF by MP3a and measurements for the Cherry building: the year of 2012 (left) and the year of 2013 (right)

5.6 Model prediction verification

We have identified two paths towards the closing of the building energy performance gap. The first path pertains to model predictive capability only. This part of the building performance gap stems from the fact that the current model or modeling procedure is inadequate to provide point predictions with which the expected performance gap evaluated over a collection of buildings is minimized.

Here, let us closely follow the definition of the performance gap, represented by the following equation:

$$d(x,y) = |x - y|, \tag{5.1}$$

where x denotes a real-valued deterministic (point) prediction and y a real-valued actual energy consumption. $d(x,y)$ represents the observed performance gap evaluated with Equation (5.1). Intuitively, the performance gap observed at one condition is equivalent to the assessment of the model prediction using absolute error function. Note that to ensure that model prediction x and actual consumption y are comparable, they should be reported using the same level of aggregation or normalization over certain temporal

period or spatial areas. Five modeling procedures are evaluated with the measurements, individually. The one that yields the minimum discrepancy with the measurement value is the winner. As in most cases no absolute winter exists that offers the minimum performance gap in any situation. We shall aggregate the results from different contexts and offer a compact statistical measure.

5.6.1 Verification of point prediction of annual cooling energy use

This section evaluates the five modeling procedures on the point predictions of the annual cooling energy consumption across six buildings. Figure 5.13 compares the predictions of MP1 against the metered energy consumption. In total, there are 12 testing points obtained from six buildings over the year of 2012 and 2013. We present the results in two ways. The left figure shows both predicted and metered values in kWh/m^2 with the 45° line as the reference. We also plot the difference between metered and predicted values over time on the right. The right one makes it easier to read the prediction errors than the left one, and also shows the changes of the prediction errors over different

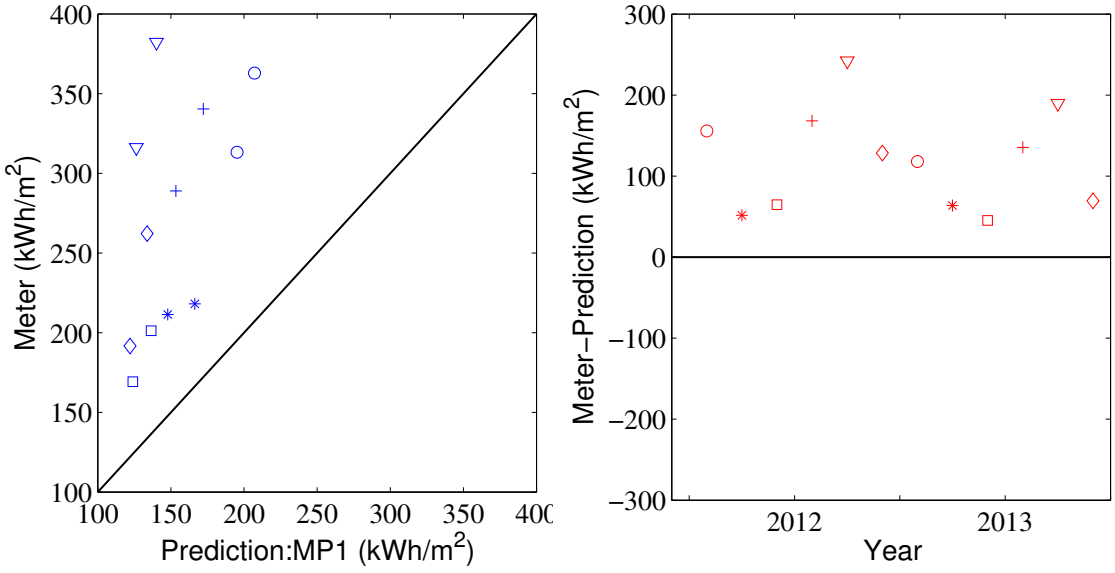


Figure 5.13 Prediction of annual cooling energy consumption by MP1 and measurements for six buildings

buildings and across different years. Different buildings are denoted with different shapes in the figures. We observe that MP1 consistently underestimates actual energy consumption, thus it provides severely biased predictions as shown in Figure 5.13.

Figure 5.14 shows the mean of the probabilistic prediction results of MP2a. Figure 5.15 shows the deterministic prediction results of MP2b. Figure 5.16 and Figure 5.17 show the mean of probabilistic prediction results of MP3a and the deterministic prediction from MP3b. For a given modeling procedure, two versions of model predictions, version a and version b, offer slightly different results, but it is not evident to visually inspect which one is the better one. If compared with MP1, MP2a or MP2b shows noticeable improvements. This indicates the importance of lighting and plug loads in prediction building cooling energy use. Specially, the lighting and plug loads derived from the actual energy use of monitored office buildings may provide better estimations on the lighting and plug loads of our campus buildings than the standardized lighting and plug loads of general office buildings. A close inspection reveals that most of the monitored buildings are indeed from the campus of universities or national labs. Thus,

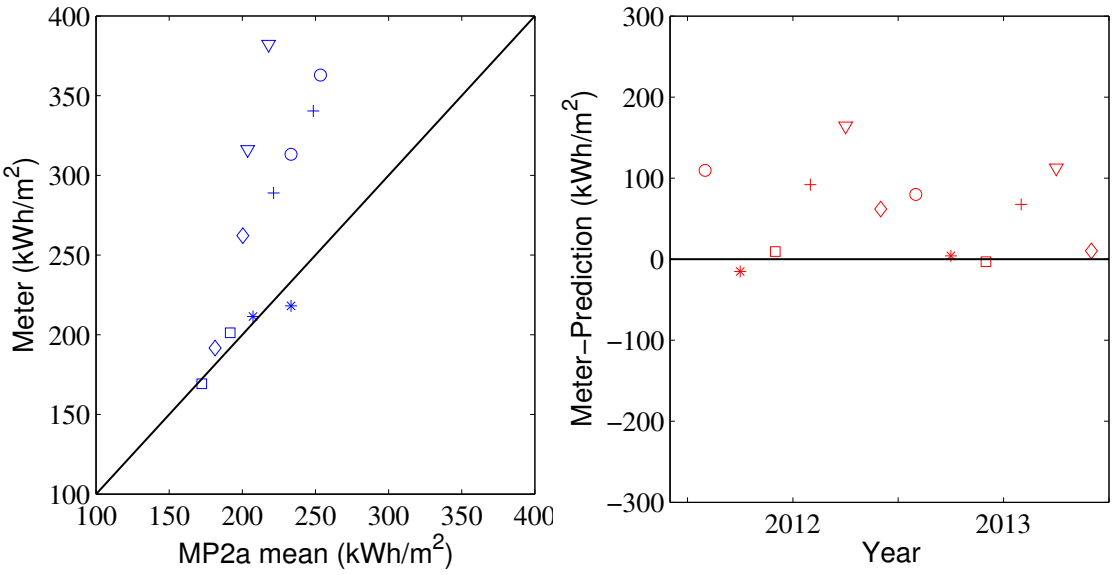


Figure 5.14 Predicted means of annual cooling energy consumption by MP2a and measurements for six buildings

the improvements of model predictions of MP2a or MP2b compared with MP1 may purely due to the good match for our campus buildings.

Even though the substantial improvements from MP1 to MP2 confirms that lighting and plug loads are important factors, both MP2a and MP2b still understates the cooling energy consumption. There must be some other important source of errors that are beyond the uncertainties of model input parameters including lighting and plug loads. Furthermore, MP3a or MP3b improves model predictions in comparison with MP2a or MP2b, showing the significance of model form uncertainty. Based on the visual inspections over different modeling procedures, we argue that model form and input parameter uncertainties are equally important in offering high-fidelity predictions.

We then continue to quantify the magnitude of the effect of parameter uncertainty and model form uncertainty. We use four statistical measures to evaluate the point predictions offered by five modeling procedures. Before aggregating across the six buildings, we first compute these measures for each individual building.

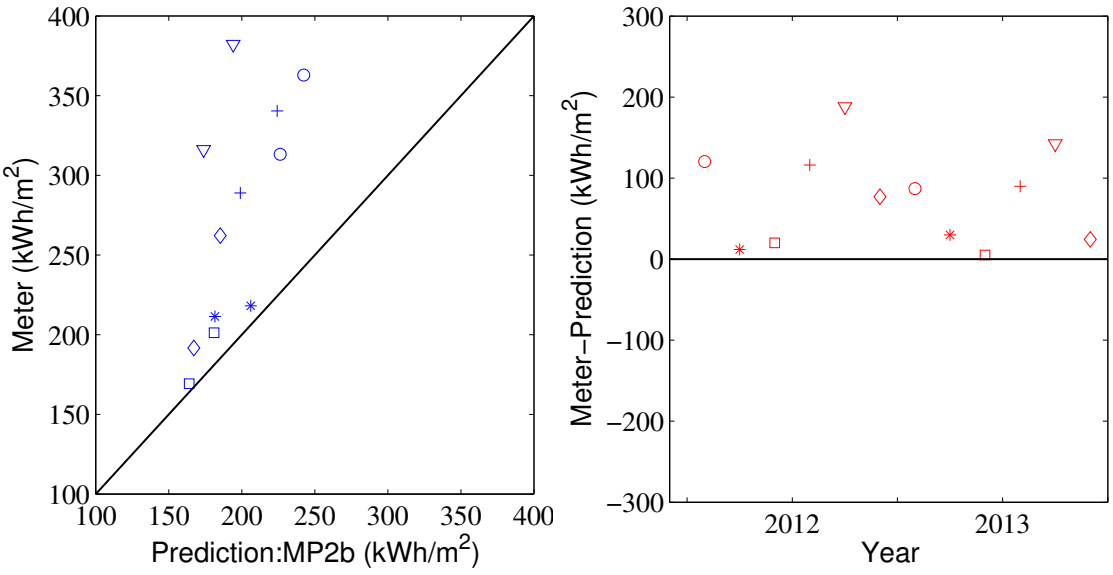


Figure 5.15 Prediction of annual cooling energy consumption by MP2b and measurements for six buildings

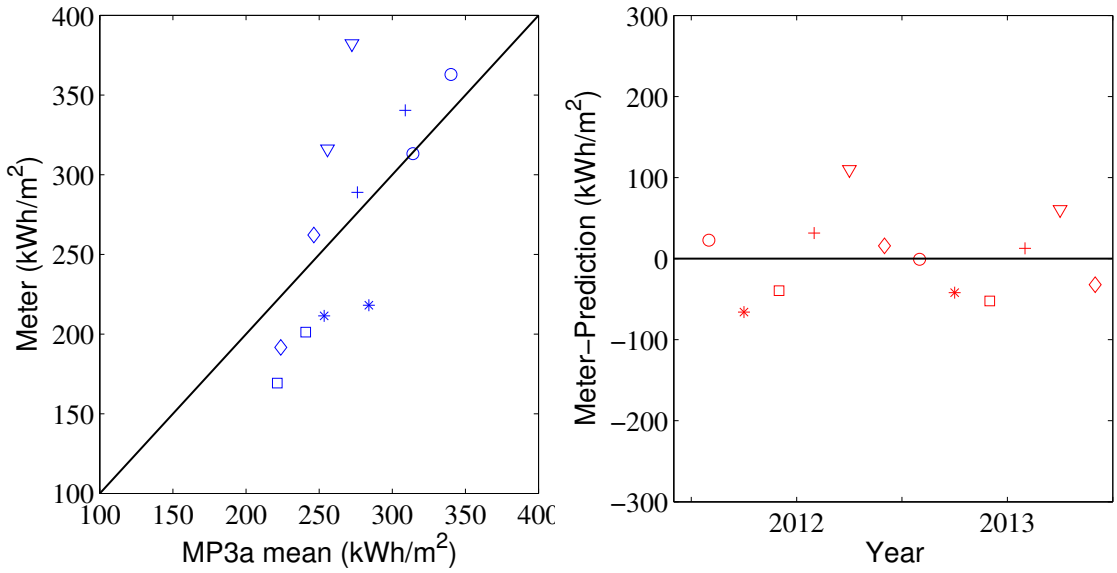


Figure 5.16 Predicted means by MP3a and measurements

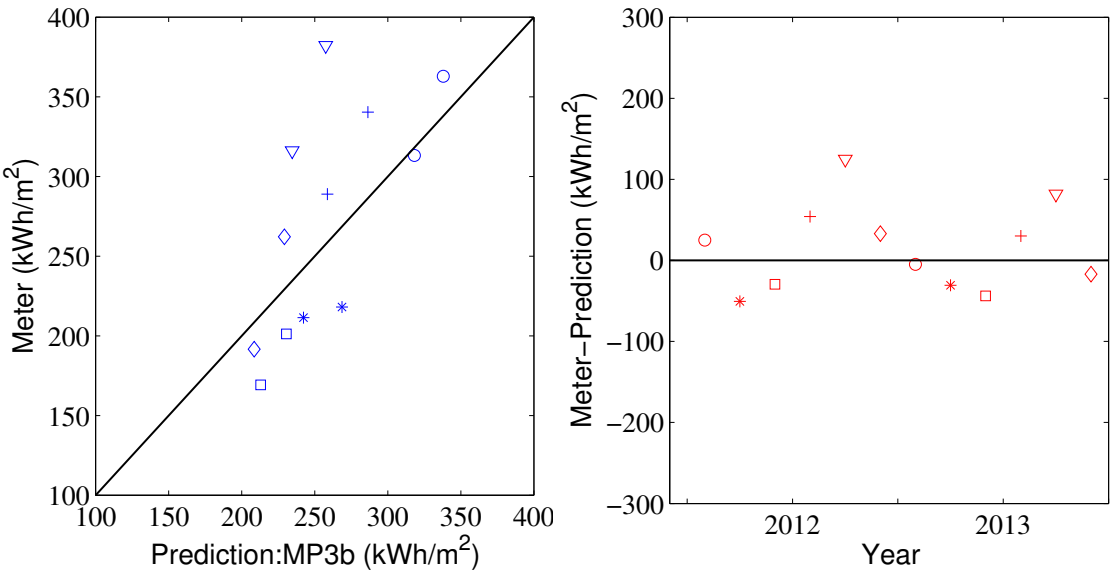


Figure 5.17 Prediction by MP3b and measurements

Table 5.3 lists the MAE, percent MAE, RMSE, and percent RMSE for each modeling procedure. Let us take the Cherry building for example. MP1 provides the baseline prediction and is also the worst performer among the five modeling procedures. It underestimates the cooling energy consumption by about 40% both in MAE and in RMSE. Taking MP1 as the baseline, MP2b and MP2a offer comparable predictions,

whose prediction error is lower than MP1 by about 10%. MP3a and MP3b offer the best predictions for this building, which reduce the prediction error from 40% to 5%. For this particular building, model form uncertainty seems to be the most important factor causing the prediction error of the standard modeling procedure, i.e. MP1. Similar findings are found in the COC, CRB, and MRDC. For the other buildings, MP2a and MP2b perform better than the other modeling procedures.

In general, the prediction accuracy of the five modeling procedures substantially varies from building to building. Figure 5.18 and Figure 5.19 use box plots to show the variation of the statistical measures over the six buildings. It becomes evident that the variation in the prediction error of MP1, MP2a and MP2b is relatively larger than that of MP3a and MP3b. For example, the MAE from MP2a may be as small as 10kWh/m^2 or as large as 150kWh/m^2 across the six buildings and over two years. In contrast, MP3a and MP3b have smaller variation in both MAE and RMSE over the six buildings, with the medians much smaller than the others. This means not only MP3a and MP3b overall perform better than the other modeling procedures, but also they are more stable since the variations among different buildings are smaller.

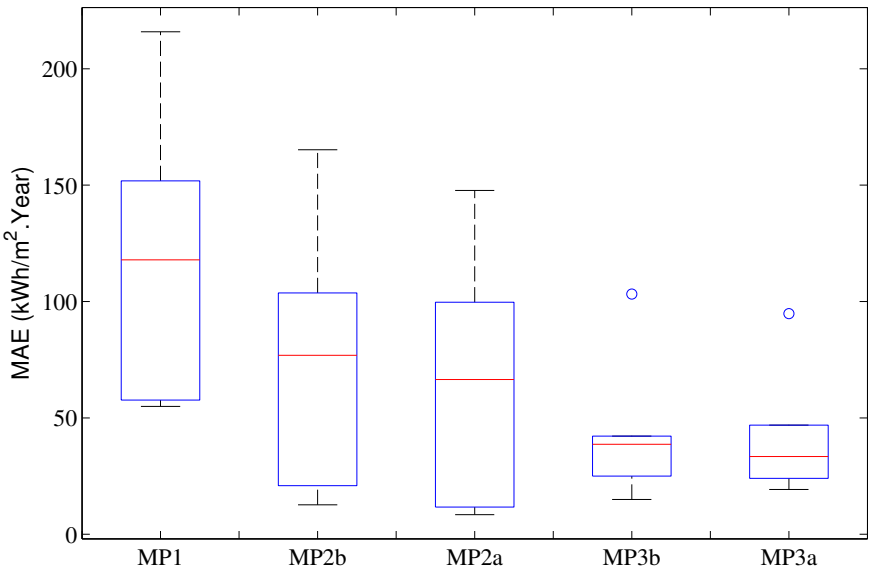


Figure 5.18 The variation of MAE on annual energy prediction among six buildings

Table 5.3 Evaluating annual cooling energy predictions of MP1, MP2b, MP2a, MP3b, and MP3a for each building using measurements in the year of 2012 and 2013

Models	MAE ($kWh/m^2.year$)	Percent MAE (%)	RMSE ($kWh/m^2.year$)	Percent RMSE (%)
Cherry				
MP1	136.80	40.46	138.10	40.85
MP2b	103.71	30.68	105.06	31.07
MP2a	99.71	29.49	95.84	28.35
MP3b	14.95	4.42	17.96	5.31
MP3a	19.27	5.70	16.12	4.77
COC				
MP1	215.89	61.82	217.47	62.27
MP2b	165.20	47.30	166.78	47.76
MP2a	147.72	42.30	140.76	40.31
MP3b	103.19	29.55	105.42	30.19
MP3a	94.81	27.15	88.66	25.39
CRB				
MP1	98.97	43.62	103.27	45.51
MP2b	50.69	22.34	57.10	25.16
MP2a	46.58	20.53	44.29	19.52
MP3b	24.98	11.01	26.23	11.56
MP3a	24.02	10.58	25.32	11.16
Health Center				
MP1	57.67	26.85	57.98	26.99
MP2b	20.89	9.73	22.73	10.58
MP2a	8.41	3.92	11.16	5.19
MP3b	40.69	18.94	41.87	19.49
MP3a	46.86	21.82	55.29	25.74
Hinman				
MP1	54.96	29.67	55.81	30.13
MP2b	12.64	6.82	14.68	7.93
MP2a	11.66	6.29	7.00	3.78
MP3b	36.64	19.78	37.34	20.16
MP3a	36.92	19.93	46.31	25.00
MRDC				
MP1	151.84	48.25	152.72	48.53
MP2b	103.09	32.76	103.93	33.02
MP2a	86.44	27.47	80.65	25.63
MP3b	42.16	13.40	43.82	13.92
MP3a	29.89	9.50	24.10	7.66

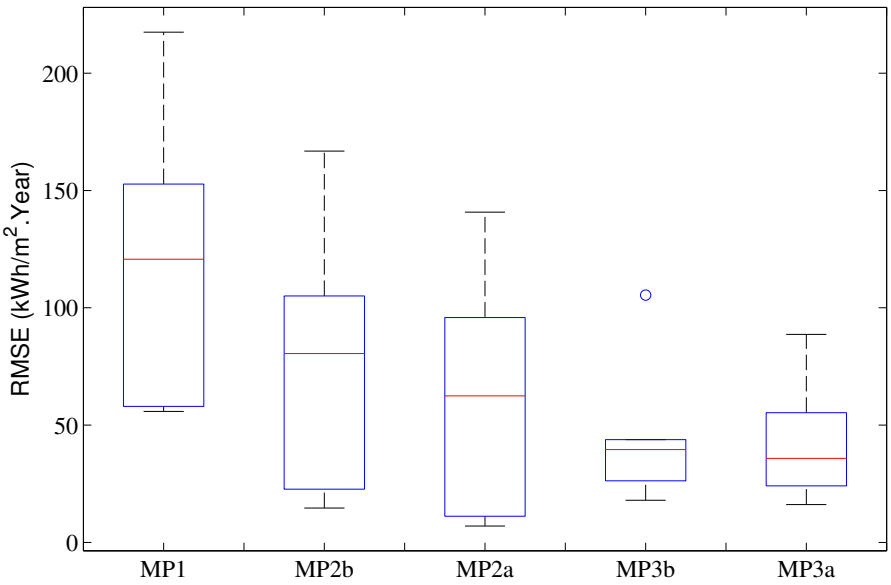


Figure 5.19 The variation of RMSE on annual energy prediction among six building

Table 5.4 lists the overall performance evaluated by the four statistical measures averaged over the six buildings. It shows that using the UQ results of the lighting and plug loads reduces the MAE from 44% (MP1) to 28% (MP2b), and reduces the RMSE from 49% to 35%. Quantifying the model form uncertainty further reduces the MAE from 28% (MP2b) to 16% (MP3b), and reduces the RMSE from 35% to 20%. It is also interesting to note that point predictions generalized from MP2a and MP3a by choosing the mean of the Monte Carlo simulations are better than the results of deterministic simulation using the means of uncertain parameters. The results show that MP2a performs better than MP2b; MP3a performs better than MP3b evaluated by the four statistical measures. For example, the percent RMSE of the MP2a is 29% that is considerably lower than that of the MP2b (35%). This indicates the impact of non-linear or interaction effects among the input parameters of the building energy models. Therefore, even for point predictions, which implicitly refer to the means, are dominant in guiding design decisions in current practice, it is worthwhile to conduct uncertainty analysis and provide the means from the distributions of the simulation results.

Table 5.4 Evaluating annual energy predictions of MP1, MP2b, MP2a, MP3b, and MP3a using measurements of six buildings

Modeling procedures	**MAE (kWh/m^2. year)**	**Percent MAE (%)**	**RMSE (kWh/m^2. year)**	**Percent RMSE (%)**
MP1	119.36	43.96	133.44	49.15
MP2b	76.04	28.01	94.56	34.83
MP2a	66.75	24.59	79.20	29.17
MP3b	43.77	16.12	53.53	19.72
MP3a	41.96	15.46	49.23	18.13

5.6.2 Verification of point prediction of monthly cooling energy use

We also evaluate the modeling procedures on monthly cooling energy prediction and show the results from Figure 5.20 to Figure 5.24. The overall findings are similar to what we have presented in the annual energy predictions. We observe patterns in the errors of MP2a and MP3b over 12 months. It seems the building energy models are more likely to underestimate the cooling energy consumption during the winter season. Due to lack of data, we could not give arguable reasons for this. One possible reason is the ignorance of spatial variability in lighting and plug loads within a building. In winter, the building has cooling demand only in some zones, whose internal loads are high. These zones may be the computer labs in the core areas of the building. These zones typically have higher plug loads than the average of the whole building. Using the averaged plug loads will thus underestimate the cooling energy use by these areas. In summer months, Figure 5.23 however, does not show systematic bias in model predictions.

Table 5.5 lists the overall performance of monthly predictions. The findings are the same as the annual case. MP3a yields the best point predictions, which sheds the prediction errors of MP1 by as much as 50%. In summary, we conclude that our work has substantially reduced the expected building energy performance gap, i.e. the average performance gap across different buildings.

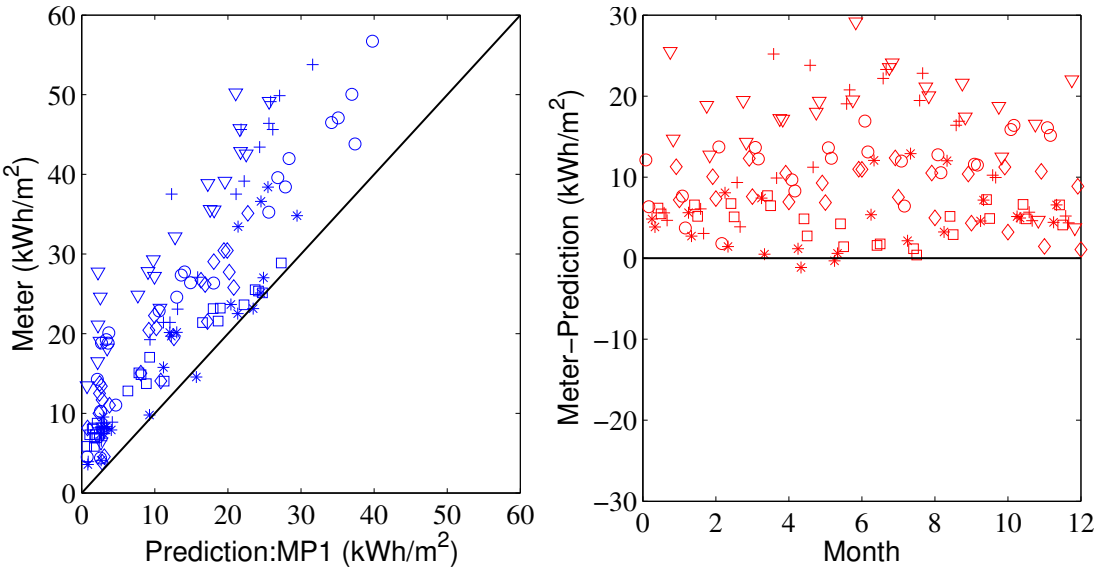


Figure 5.20 MP1 predictions and measurements of monthly cooling energy use

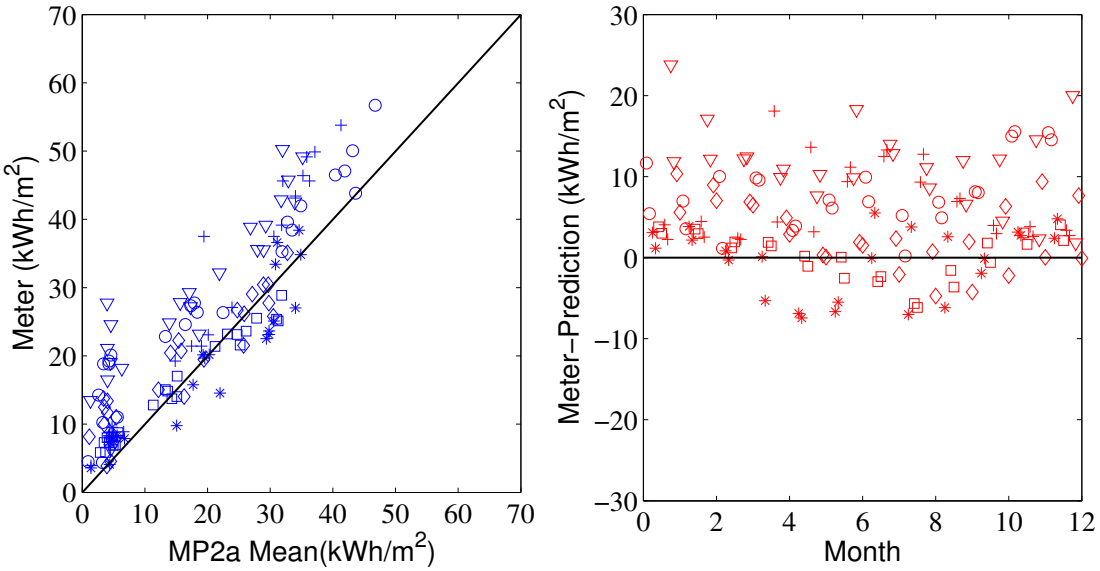


Figure 5.21 Predicted means of MP2a and measurements of monthly cooling energy use

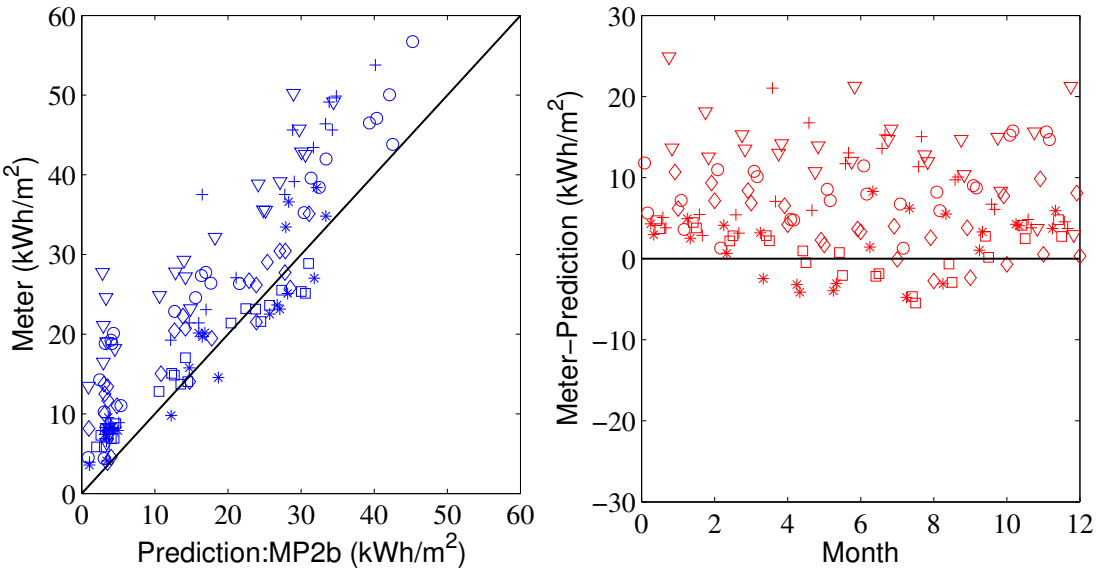


Figure 5.22 MP2b predictions and measurements of monthly cooling energy use

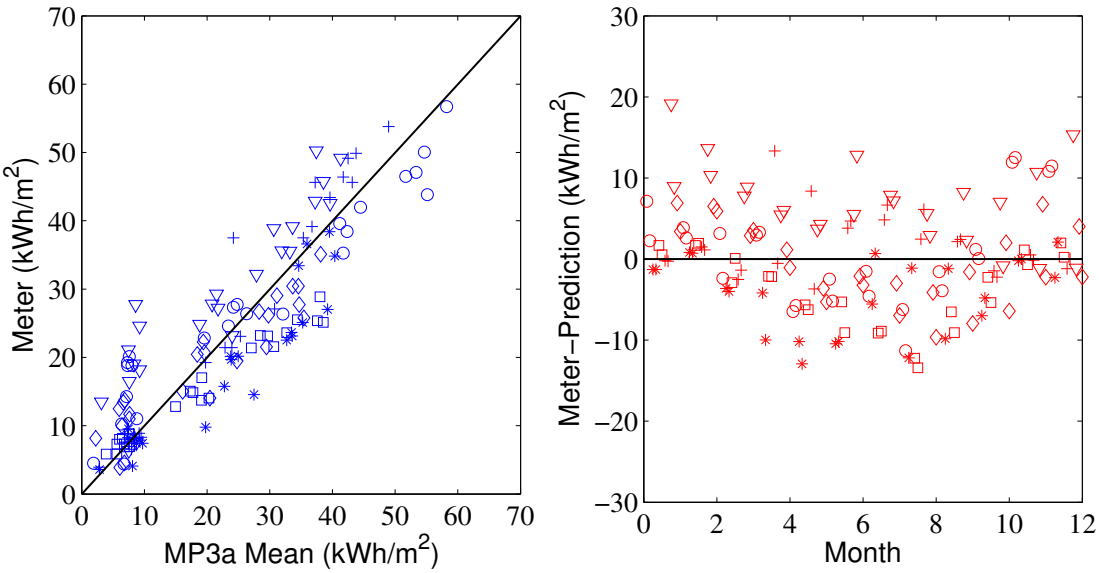


Figure 5.23 Predicted means of MP3a and measurements of monthly cooling energy use

Table 5.5 Evaluating monthly energy predictions of MP1, MP2b, MP2a, MP3b, and MP3a using measurements of six buildings

	MAE (kWh/m2. Mo)	**Percent MAE (%)**	**RMSE (kWh/m2. Mo)**	**Percent RMSE (%)**
MP1	9.97	44.05	11.96	52.88
MP2b	7.04	31.13	8.77	38.74
MP2a	6.40	28.29	7.70	34.05
MP3b	4.94	21.84	6.43	28.40
MP3a	4.79	21.18	6.26	27.66

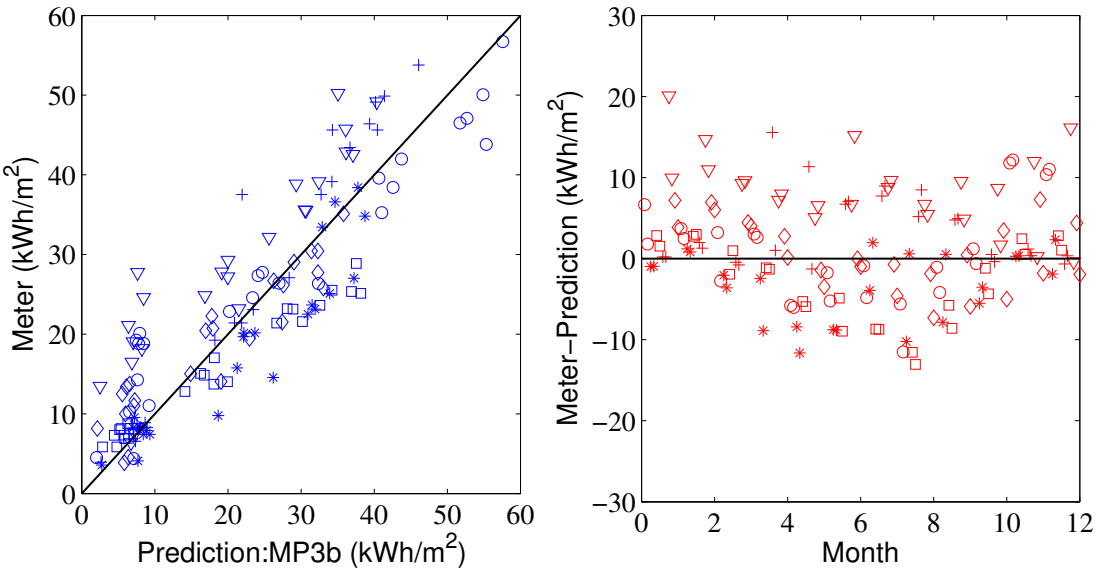


Figure 5.24 MP3b predictions and measurements of monthly cooling energy use

5.6.3 Probabilistic prediction verification using PIT method

Figure 5.25 shows the probabilistic predictions of MP2a on the annual energy consumption. The prediction intervals of MP2a cover only 7 out of the 12 measurement points. Although there are several measurement points that fall close to the means of the probabilistic predictions, most measurement data are higher than the prediction means.

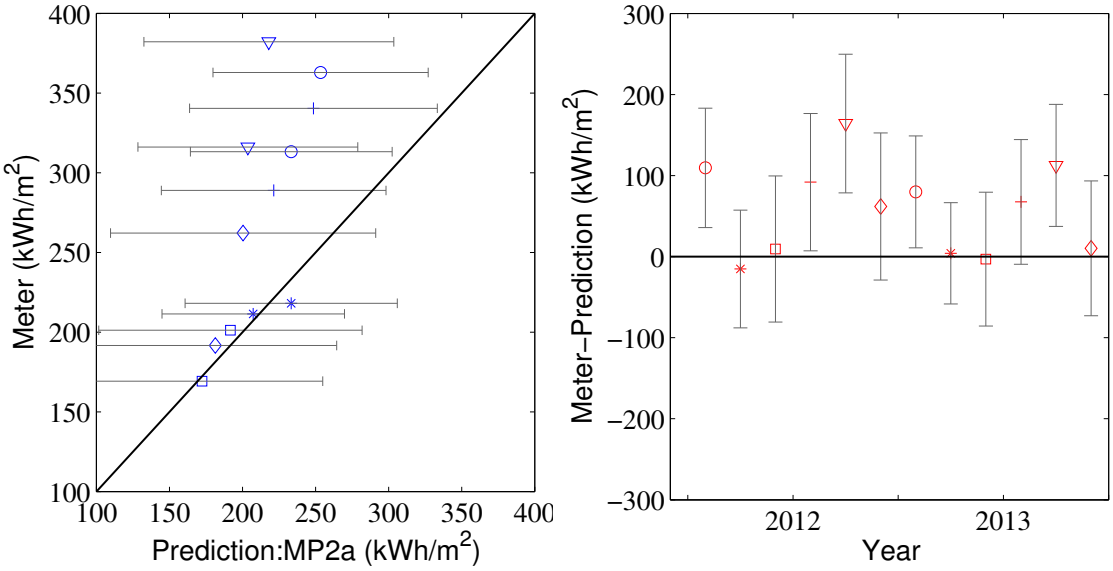


Figure 5.25 Predictions by MP2a and measurements of annual cooling energy use

Figure 5.26 shows the probabilistic predictions of MP3a whose prediction intervals cover 11 out of the 12 measurement points. Moreover, the real-valued measurements fall on the right halves of the predicted intervals for some cases and on the left halves of the predicted interval for other cases, which is required for the probabilistic predictions to be verified.

Figure 5.27 shows the CDF of the probability integral transform (PIT) evaluated over 6 buildings. As we have demonstrated using the simulated example in Chapter 4, the PIT should be uniformly distributed from 0 to 1 if the modeling procedure is statistically verified. The figure shows that the probabilistic predictions from MP2a are biased. In particular, MP2a underestimates the actual cooling energy consumption. In contrast, the CDF of the PIT of the MP3a seems more uniformly distribution from 0 to1. To be rigorous, we also employ the K-S hypothesis test. The null hypothesis is that the PIT is uniformly distribution from 0 to 1. The K-S test rejects the null hypothesis for the MP2a with p-value of 0.001, whereas the K-S test fails to reject the null hypothesis for the MP3a at a significance level of 0.05. Therefore, we shall accept that the PIT from MP3a is uniformly distributed from 0 to 1. Hence, the probabilistic predictions using MP3a are

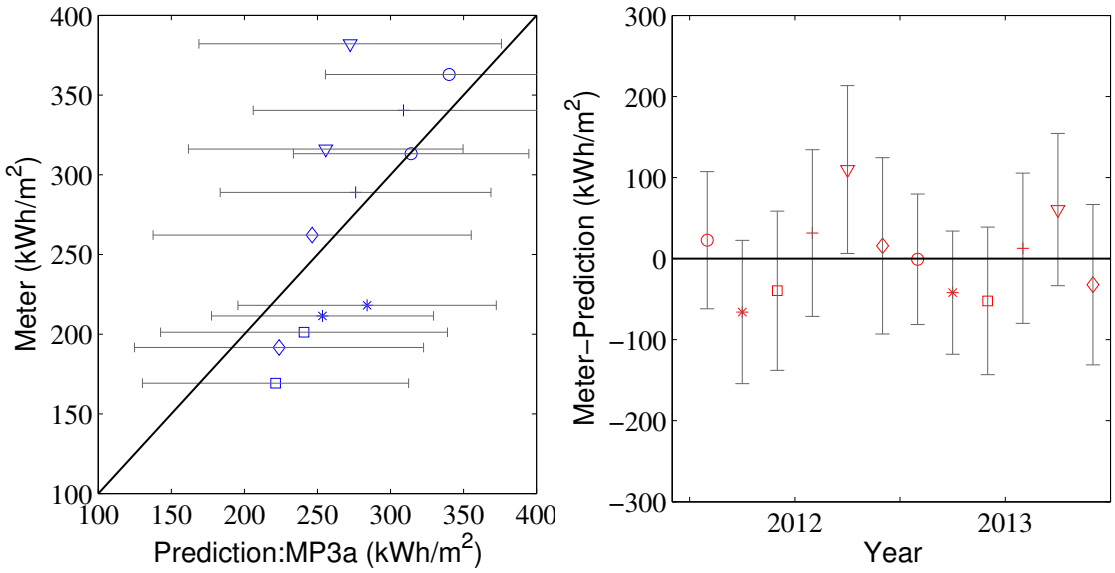


Figure 5.26 Predictions by MP3a and measurements of annual cooling energy use

statistically verified, showing the importance of model form uncertainty quantification.

Figure 5.28 shows the CDF of the PIT from monthly probabilistic predictions. Each CDF contains 144 data points. It shows that the CDF of MP3a is closer to the uniform distribution than the MP2a. However, the K-S test rejects the null hypothesis with a p-value of 0.000 for MP2a and a p-value of 0.004 for MP3a.

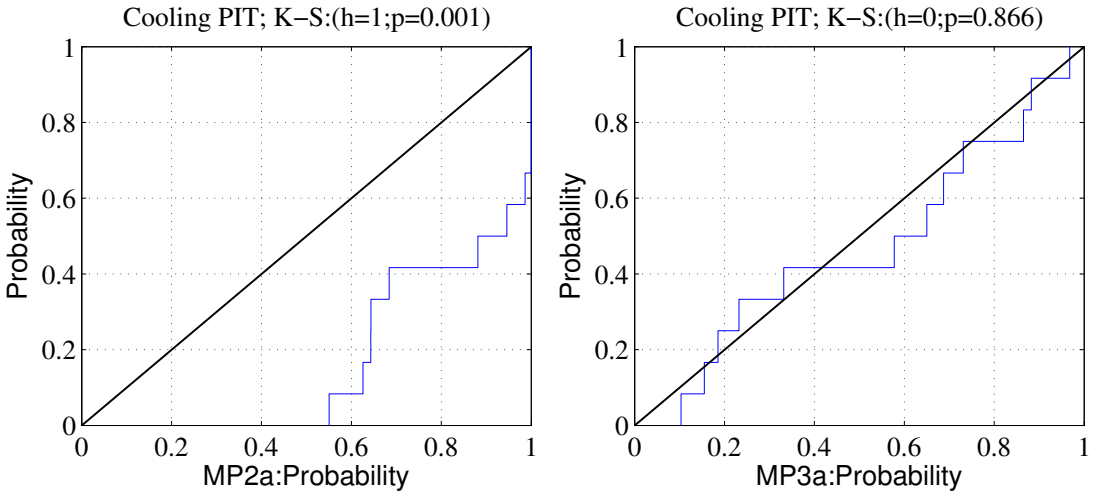


Figure 5.27 PIT CDF and K-S test for MP2a MP3a probabilistic predictions of annual energy consumption of six buildings in 2012 and 2013 (12 data points in total)

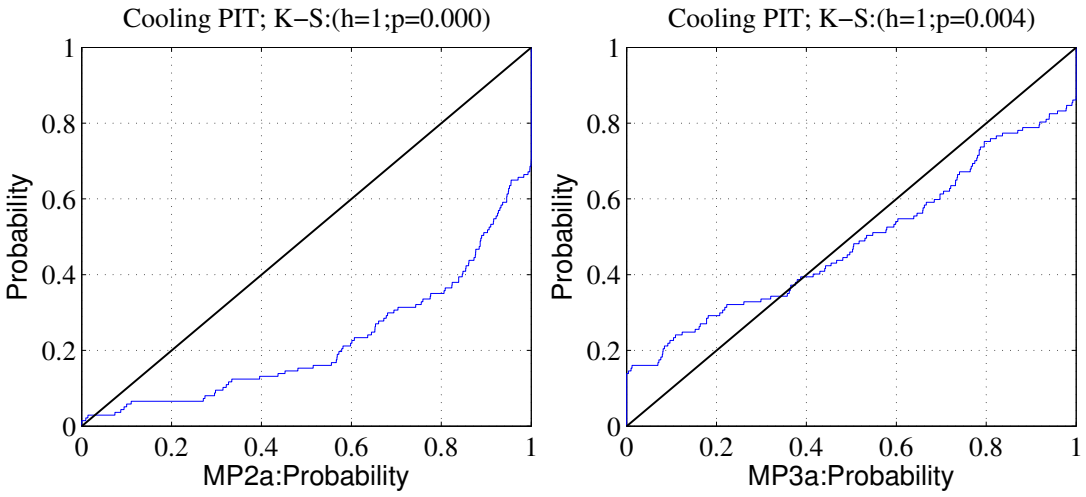


Figure 5.28 PIT CDF and K-S test for MP2a and MP3a probabilistic predictions of monthly energy consumption of six buildings in 2012 and 2013 (144 data points in total)

5.6.4 Probabilistic prediction verification using CRPS method

In addition to the PIT method that provides visual inspections of the probabilistic prediction, we then compute the continuous ranked probability scores (CRPS) for MP2a and MP3a for both monthly and annual probabilistic predictions to quantitatively evaluate the two modeling procedures. The CRPS results are given in Table **5**.**6**. It shows that the MP3a offers better probabilistic predictions than the MP2a. Specifically, MP3a reduces the CRPS value, which is equivalent to the MAE in the point prediction case, by about 50% on annual probabilistic predictions, and by about 25% on monthly probabilistic predictions. These results further confirm the importance of model form uncertainty in uncertainty analysis at whole-building level.

5.7 Sensitivity analysis

As for the monthly cooling energy use, predictions of MP3a do not match with the measurements so well as the annual cooling energy use. There seems to be a need to better quantify the uncertainty of either model parameters or model forms. However, the effects of further improvements from MP3a are expected to be marginal as shown by Figure 5.28. Further improvements of monthly probabilistic predictions are considered as the further work of this thesis. As for annual cooling energy consumption, since the MP3a offers statistically verified probabilistic predictions based on our current measurement data, there seems to be no urgent need to further enhance the model predictions through better quantification of uncertainties before new data are obtained. Without new verification data, we shall consider MP3a able to provide probabilistic

Table 5.6 Continuous ranked probability score for evaluating the probabilistic predictions of monthly and annual energy use by MP2a and MP3a

Monthly CRPS (kWh/m^2. Mo)		**Annual CRPS (kWh/m^2. Year)**	
MP2a	MP3a	MP2a	MP3a
4.88	3.68	52.77	27.83

predictions of the annual district cooling energy use for similar types of buildings. If MP3a is to predict annual cooling energy consumption for a new building, its probabilistic predictions shall be directly used to guide design decisions.

Our UQ repository including both parameter and model form uncertainty quantification has shown to decrease the performance gap of annual total cooling energy use from 44% (i.e., the current standard modeling procedure MP1) to 15% (i.e., MP3a). Recall the two major types of uncertainty described in Chapter 3, i.e., model form uncertainty and input parameter uncertainty. This significant reduction thus comes from the fact that the distributions of the uncertain parameters in the UQ repository may better describe the characteristics of our campus buildings and that the model form uncertainty quantification at different subsystems may fundamentally enhance the fidelity of the complete building energy model. Likewise, the remaining 15% of MP3a predictions can also be decomposed into input parameter uncertainty and model form uncertainty. Further reducing the uncertainty, however, would not be the primary concern of modelers who have already issued statistically verified probabilistic predictions. Speaking differently, the main part of 15% is expected to be irreducible uncertainty from the modelers' perspectives. Therefore, the focus is to shift from quantification to management of uncertainties. This section employs sensitivity analysis with the aim to offer advices on how to reduce the uncertainty of model predictions by determining the most influential uncertain parameters.

From Figure 5.29 to Figure 5.34, we plot the SI of uncertain parameters for the annual cooling energy consumption for six buildings. The SA is conducted with the two-step method we developed in Chapter 3. The R^2 values are typically higher than 0.95 after we consider the first-order main effect. Moreover, the ranking of the uncertainty parameters are quite consistent across the six buildings. Specifically, all building models rank lighting and plug load peak use and schedule uncertainty as the top two uncertain

contributors. The lighting and plug load related uncertainties explain about 80% of the total uncertainty in annual cooling energy predictions. The remaining 20% uncertainty is attributed to different parameter in different cases, including weather, convective heat transfer coefficients, and effective leakage area. Based on the SA, uncertainty management should be focused on the uncertainty in lighting and plug loads to reduce the prediction uncertainty stemming from uncertainty parameters.

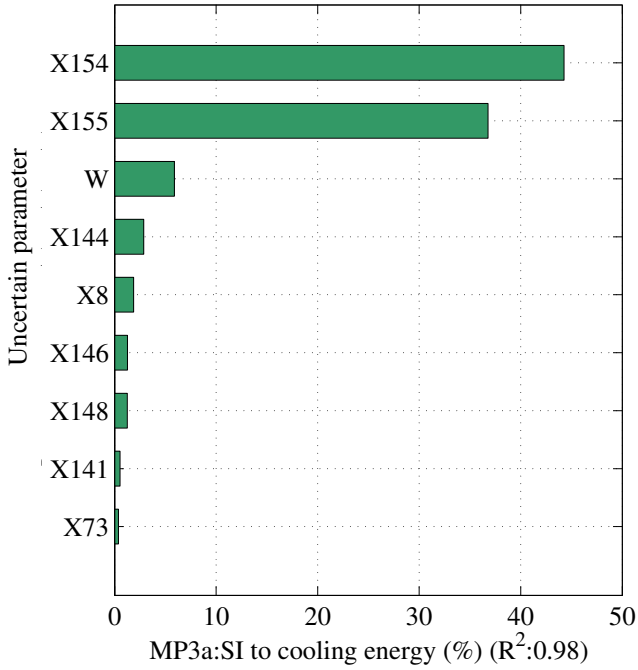


Figure 5.29 MRDC building (X154: L&PL peak density; X155: L&PL schedules; W: weather)

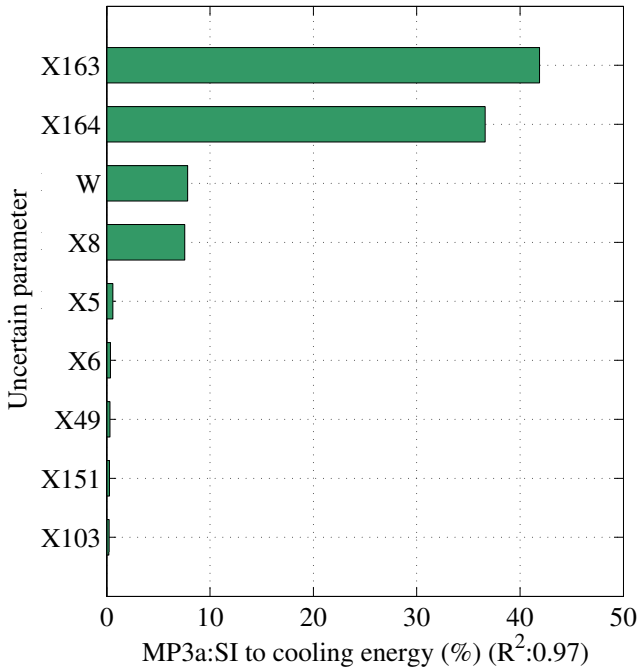


Figure 5.30 Cherry building (X163: L&PL peak density; X164: L&PL schedules; W: weather; X8: effective leakage area)

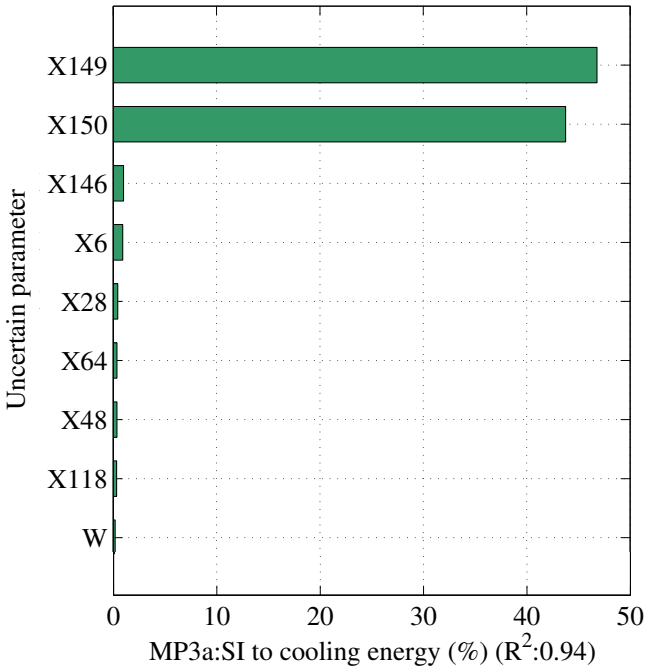


Figure 5.31 Hinman building (X149: L&PL peak density; X150: L&PL schedules)

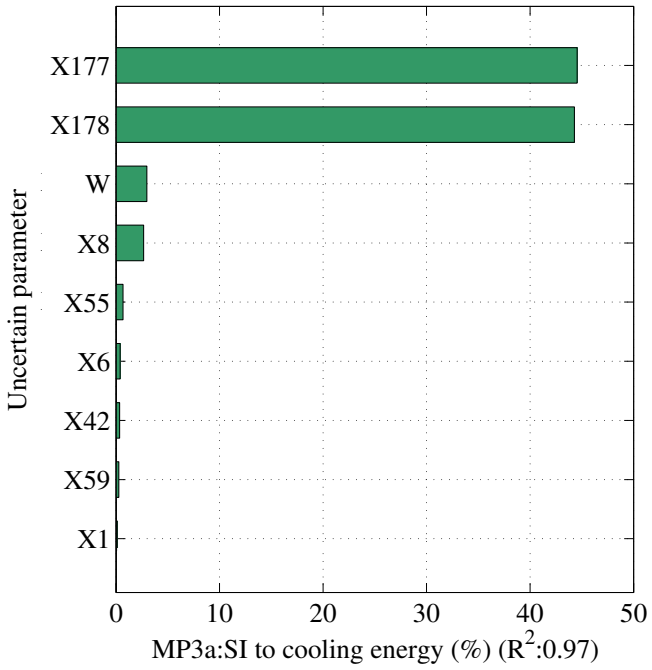


Figure 5.32 CRB building (X177: L&PL peak density; X178: L&PL schedules; W: weather; X8: effective leakage area)

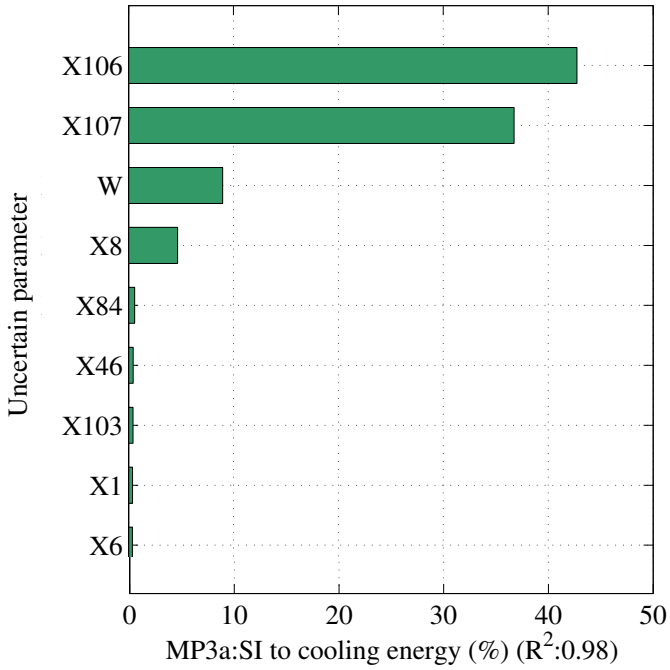


Figure 5.33 SA of cooling energy consumption for Health Center building (X106: L&PL peak density; X107: L&PL schedules; W: weather; X8: effective leakage area)

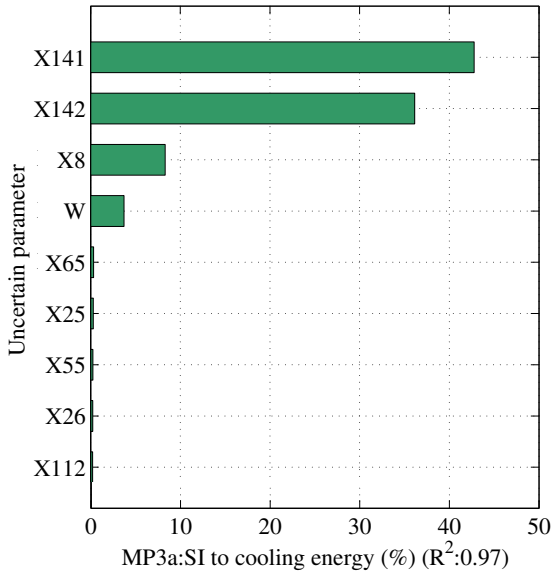


Figure 5.34 SA of cooling energy consumption for COC building (X177: L&PL peak density; X178: L&PL schedules; W: weather; X8: effective leakage area)

5.8 Summary and discussion

This chapter demonstrates the methodology developed in previous chapters and reaches several findings based on six case buildings on the Georgia Tech campus.

1. Model from and parameter uncertainty are two major types of uncertainty affecting the prediction capability of building performance simulations. Our UQ repository, which quantified both types of uncertainty, is shown to substantially enhance the prediction accuracy. Taking the annual district cooling energy use as an example, the mean absolute errors decrease from 44% to 15%. The remaining 15% of the errors is likely to be mainly caused by inherent uncertainties that are unlikely to be reducible through modelers' endeavors. Speaking differently, approximately $\pm 15\%$ discrepancy between predicted and measured energy use of district cooling energy is achievable by improving the prediction capability of building performance simulations.

2. The considerable amount of error reduction, e.g., from 44% to 15%, is the joint effect of model form and parameter uncertainty quantification. Using the

EnergyPlus V7.0 and the related modeling recommendations as the references, model form uncertainty and parameter uncertainty have comparable significance toward improving prediction accuracy. The PIT results show that statistically verified probabilistic predictions on annual district cooling can be issued if and only if model form uncertainty has been adequately dealt with. This also means that the ignorance of model form uncertainty under current modeling standards is questionable in model-based applications, for example, estimating the values of parameters through data-driven calibration methods for retrofit recommendations purpose.

3. Our UQ repository has also decreased the monthly prediction errors from 44% to 21% measured by the MAE. Although the improvement is substantial, the K-S test of the PIT distribution rejects the null hypothesis of uniform distribution. This may caused by other ignored sources of uncertainties, such as the HVAC systems controls. This may also result from the deficiency in representing the uncertainties by the current UQ repository. Based on the patterns of monthly prediction discrepancy over a year, we suspect the uncertainty in some parameters, e.g. lighting and plug loads, which are currently quantified at the whole-building scale, had better be refined to the level of thermal zones.

In addition to the findings from the six case studies, there are several issues worthy of extended discussion.

5.8.1 Data quality

The concept of "fitness for use" is widely adopted in assessing the quality of products in general. "Data quality" is thus defined as data that are fit for use by data consumers (Wang and Strong 1996). Prediction verification uses data as benchmarks with which model predictions are compared and evaluated. From this particular perspective, some attributes associated with data quality include data accuracy, completeness, and variety.

Data accuracy is mainly concerned with whether the measurement errors are controlled to such a level that can be safely ignored without affecting subsequent analysis. This typically involves the quantification of measurement errors either of random or systematic and followed by impact assessment, with detailed methods referred to Coleman et al. (2009). Sometimes it is difficult to predetermine whether certain magnitude of measurement errors have detrimental effect based on simple engineering judgment. If so, measurement errors will have to be explicitly dealt with as another source of uncertainty over the entire verification process. Readers are referred to ASME (2009) for implementation details. Quantification of measurement errors requires additional information just like what we have quantified the other sources of uncertainties. Here, it may include uncertainty quantification of each individual component, such as the calibration accuracy of sensors, assembled into the overall experimental apparatus to derive the resultant uncertainty of the final quantity of interest. For example, the chilled water energy is derived from three measured variables as shown in Figure 5.7. To quantify the uncertainty in measuring the chilled water energy, we need to quantify the uncertainty for each measured variable and then calculate the uncertainty through uncertainty propagation method for example (Coleman et al 2009). In our work, we implicitly assume that the data of monthly or annual cooling energy use are of high accuracy. This assumption may hold because of the reputation of the installation company, the maturity of the measurement technology, and the aggregation use of the data. Nevertheless, systematical measurement error analysis can be done if required information is obtained. It is suggested that building performance-monitoring systems have standard procedures of accuracy assessment and reporting, which unfortunately is not the case at present.

In large data sets, missing data have always been commonplace for a variety of reasons. Missing data can be handled either by removing them from the verification data set, or by estimating them from related data through interpolating nearby data points for

example. The latter is recommended if the number of missing points is small and the estimates are reliable. When many successive missing points occur, removing that portion of the data may be the only choice. For model verification use of data per se, if data are missing at random and the data sets is large, the overall evaluation statistics, e.g., MBE, is likely to be relatively unaffected by the missing data (Jolliffe and Stephenson 2003).

In terms of data variety, it is always desirable for a data set to include simultaneous measurements of different types of variables, such as cooling and heat separately submetered energy use, lighting and plug loads, and indoor temperature conditions. When different types of variables can be simultaneously monitored over a relatively long period, model verification can be performed in a much more comprehensive manner than we do on our buildings. Moreover, comparing the accuracy among different types of variables may also lead to hints about the sources of prediction discrepancy. Regarding types of variables, a hierarchy of building energy performance metric has been developed in the literature (Barley et al. 2005) to improve the consistence in reporting the results.

5.8.2 HVAC systems uncertainty quantification

The current UQ repository has emphasized on thermal demand related uncertainties, yet has not completed the uncertainty quantification involving HVAC systems control and operations. Because the case buildings use district-cooling systems that do not contain the major HVAC energy consumer such as chillers, HVAC systems uncertainties thus do not play a major role in predicting the energy performance of our campus buildings. This explains the current UQ repository can issue statistically verified probabilistic predicted shown by the PIT method even though it has not included HVAC systems uncertainties. In fact, uncertainties in HVAC systems could be the other important component causing the performance gap for buildings with chilled water

systems. Although the campus buildings provide favorable results, which are unlikely to be so for buildings with complete HVAC systems, we consider these buildings are appropriate to examine the effect of our current UQ repository on thermal demand predictions relatively in isolation of chilled water systems. With the encouraging results, we should be confident in adding more types of uncertainties towards an exhaustive quantification of uncertainty for building performance simulation.

CHAPTER 6 TOWARDS INFORMATIVE DECISIONS: PROBABILISTIC PREDICTION FOR HVAC SYSTEMS SIZING

Previous chapters show how high-fidelity probabilistic predictions are obtained through comprehensive quantification of uncertainties. Hence, we claim the building energy performance gap can largely decrease by improving our predictions. As shown in previous chapters, probabilistic predictions enable the expression of uncertainty for estimating the actual building performance, thus providing more information and encouraging honest predictions than the deterministic framework. In order to procure energy efficient buildings, better information gained from high-fidelity predictions has to be transformed into actual savings of energy through informative decisions. This chapter explores how probabilistic predictions can add value to the traditional method of HVAC systems sizing.

6.1 Introduction

In current practice, HVAC systems are sized based on standardized procedures that were mostly developed by ASHRAE. The standard approach only implicitly deals with uncertainty in peak system demand through the selection of an appropriate design day and the choice of a safety factor. Although this method works satisfactorily in most cases, it offers no transparency of the probability that a system design does not meet a required level of performance. When a system designer wants to track the risk associated with an undersized system, the standard method does not supply enough information. The opposite, i.e. avoiding that the system is needlessly oversized deserves even more attention given the fact that current practice of "defensive sizing" leads to oversized systems which leads to wasted capital investment and systems that operate far away from the optimum efficiency loads. This chapter explores a new framework to guide the use of probabilistic prediction and sensitivity analysis (SA) in HVAC system sizing.

Probabilistic prediction will replace the safety factor with quantified margins based on comprehensive quantification of different sources of uncertainty. A probabilistic-based SA is then used to identify the important individual factors or groups of factors that contribute to uncertainty, providing means of risk management by applying better quality assurance methods or negotiating performance contracts.

6.2 Literature review

As a major energy consumer, the HVAC system is designed to fulfill the functional requirements of maintaining thermal comfort and control relative humidity given a schedule of use and loads of the building zones. The approach to HVAC system design has been largely reduced to prescriptive procedures developed by ASHRAE (2009a). Heating and cooling load prediction is fundamental for sizing boilers, chillers, coils, piping, ductwork, terminal devices, and every other components of the HVAC system. Therefore, load prediction significantly affects the choice of components and their sizing, which thereby determines system first cost, operational cost, and energy consumption. Indirectly, a verification of how well a chosen HVAC system meets thermal (and acoustic) comfort requirements during certain extreme loads or weather events over time is also dependent on our ability to make adequate load predictions that reflect expected variability.

The current load calculation method has been anchored in the ASHRAE Handbook of Fundamentals for decades. In spite of recognized inadequacies, a shift towards reliability engineering (O'Connor and Kleyner 2011) as witnessed in other engineering fields, has not happened yet in the building systems domain. To make up for this, a safety factor is still prevalent in HVAC sizing to manage uncertainty, or rather to avoid a system that is undersized to perform adequately in all potential (but unspecified) circumstances. The safety factor is deemed to render the resulting system robust enough to deal with unspecified weather and load conditions. Since the choice of an adequate

safety factor is highly dependent on the designers' experience, the resulting variation in the robustness of the system design could be large. Moreover, modern buildings are becoming progressively complex, relying on the orchestration of local and global controls and energy saving technologies. The "defensive" and blind application of a safety factor can easily lead to excessive oversizing, considering that the design engineers are prone to minimize their professional risk. Oversizing practices of air-conditioning systems were found to be extensive in practice. For example, it was reported that over 40% of the rooftop units were oversized by more than 25% (Felts and Bailey 2000). In another study a method was proposed to quantify the oversizing ratio based on measured cycling rate and found that it was not uncommon that systems were oversized by as much as 100% (Djunaedy et al. 2011). These studies support the general consensus that oversizing may be happening across the board, rooted implicitly in the current methods. In this paper we explore the cause and effect and propose methods for HVAC sizing that express a safety factor that is based on uncertainty analysis.

6.3 Proposed HVAC system sizing framework

The traditional method of HVAC system sizing is described in the Chapter 14 to 19 of the 2009 ASHRAE Handbook of Fundamentals (ASHRAE 2009b). Hourly zone-by-zone peak heating and cooling demands are determined as the main quantity of interest (QOI) that directly affects system sizing.

It has been well recognized that the actual heating and cooling demands required to size the system are always uncertain. This is fundamentally so, because the predictions of system demands relate to the future as-built state and it is impossible to predict the future with precision and certainty. Among different inputs, the traditional method gives special attention to the weather uncertainty, which lead to the development of the design day method. A design day consists of 24 hourly values of weather variables, representing the extreme weather conditions under which the system should be able to maintain the

indoor environment at a desired temperature set point schedule and within a relative humidity band. Design days are derived from multiple actual meteorological years (AMY), introducing a variety of approximations (ASHRAE 2009b). For example, multiple weather variables (e.g., dry- and wet-bulb temperature, solar radiation, and wind speed) that affect building cooling and heating demands are highly correlated. The developed design days, however, only take the correlation between dry-bulb and wet-bulb temperature into account, whereas the rest of important variables, such as solar radiation, are assumed to be independent. Because of the correlations, the extreme conditions of different weather variables are unlikely to occur simultaneously. The independence assumption can lead to an unrealistically severe weather day. Hence, the use of a design day overestimates the weather extremes in the HVAC sizing. We will investigate this in a case study in section 4.

After the input data are collected, the calculation procedure is done in a deterministic fashion. Many of the nominal values for inputs are given as reference values in the ASHRAE Handbooks and standards (ASHRAE 2007b). As shown in the comparison in Figure 6.1 (a), the traditional framework takes these standardized inputs to generate a point prediction, using for example the heat balance method. Although the deterministic prediction may approximate the averaged peak value of the hourly demand, the spread of the possible outcomes remains unknown. HVAC designers then apply safety factors to further safeguard the system design against unknown effects.

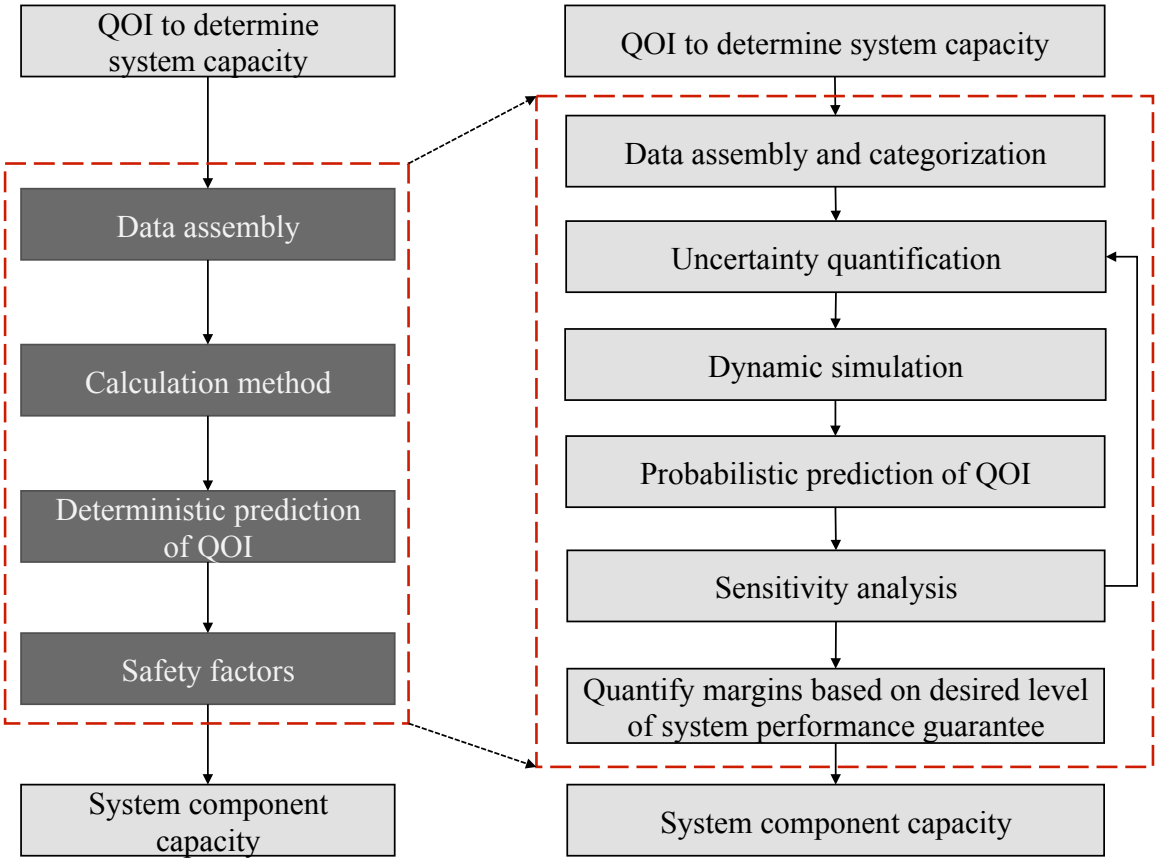
In contrast to the above method, the proposed framework (Figure 6.1 (b)) consists of the following six steps.

Step 1: Data assembly and categorization

Input data, which are also required by the traditional method, include architectural design specifications, such as building shape, outdoor weather data, indoor set-point conditions, and internal heat gains and operational schedules. Regarding the outdoor

weather data, the proposed method uses multiple actual meteorological years (AMY) which include weather variables such as temperature, humidity, wind speed, and solar radiation. The data can be freely obtained from the National Climatic Data Center (NOAA) for the majority of cities in the US. Because of climate change effects, only the recent weather years after 1982 are recommended for the use of system design (ASHRAE 2009b).

The data are categorized into two different types: (1) design parameters, and (2) uncertain parameters. Design parameters are those that designers can control, e.g., window areas. We are interested in predicting the system performance for given architectural design specifications. From this viewpoint, design parameters are those that



(a) Traditional framework (b) Proposed framework

Figure 6.1 Comparison of traditional and proposed system sizing framework

can be regarded as certain parameters in the model. The second type pertains to those non-design parameters whose values cannot be known with enough certainty in the design stage when simulation is called upon. These uncertain parameters occur in the specifications of external weather conditions, building envelops, material properties, and internal usage conditions. They can be further classified based on practical needs. For example, usage, occupancy and set point schedules can be treated as a range of "scenarios" and analyzed separately from the other types of uncertainty.

Step 2: Uncertainty quantification

It is imperative to ensure the UQ step is carefully performed because it is the basis of any further analysis. Parameter UQ, which is a difficult task, has not received enough attention in previous research. More recently, the situation has improved, and a consolidated UQ repository (Lee et al. 2013b) that is built on previous (De Wit and Augenbroe 2002; Macdonald 2002) and more recent studies (Cóstola et al. 2010; De Wit and Augenbroe 2002; Desogus, Mura, and Ricciu 2011; Domínguez-Muñoz et al. 2010; Li, Harvey, and Kendall 2012; Wang, Mathew, and Pang 2012; Sun, Heo, et al. 2014; Lee et al. 2012) has emerged.

Besides parameter uncertainties, model form uncertainty also affects the probabilistic predictions (Oberkampf and Roy 2010). As we have proved in Chapter 5, ignoring the model form uncertainty leads to bias prediction either in deterministic or probability form. For all different types of uncertainty, probabilistic theory is the appropriate rigorous basis to characterize the uncertainty. This provides a uniform basis for the representation of uncertainty and their mathematical and computational treatment (O'Hagan and Oakley 2004).

Step 3: Dynamic simulation

The reason to use the dynamic simulation in this context is twofold: (1) system sizing requires hourly predictions of cooling and heating demand, and (2) dynamic

simulation uses a physical principles based high-fidelity model so that the total model bias, which is very difficult to quantify, is relatively small compared to the effect of parameter uncertainties. Moreover, dynamic simulation tools are already popular within the building performance community. There will hence be no steep learning curve as it is straightforward for that community to use the same simulation for system sizing.

Step 4: Probabilistic prediction of the QOI

Sample-based methods such as Monte Carlo based simulations, are widely used in UA with application examples across many engineering domains (Allaire 2009; Helton and Sallaberry 2009; Helton and Davis 2003; Janssen 2013). The technical details and implementation procedures are given in section 3.

Step 5: Sensitivity analysis

Sensitivity analysis (SA), and probabilistic SA in particular, is concerned with how the total output uncertainty can be attributed to uncertainties in individual inputs or groups of inputs. SA ranks the importance of input parameters based on their influence on the uncertainties of model outcomes.

Step 6: Quantify margins based on desired level of system performance guarantees

The UA generates a probabilistic prediction of the QOI, which is represented as a cumulative density function (CDF). Figure 6.2 shows hourly peak cooling demand as an example to illustrate the idea. Suppose the HVAC designer intends to predict the i^{th} maximum hourly cooling demand over a whole year, i.e., 8760 hours, which we refer to as peak cooling demand D, to make the notation simple. The probabilistic prediction of D, which pertains to the future, depicts the uncertainty characterized by the CDF curve where we denote the expected value of D as $E(D)$. Given the CDF, the stake holder (e.g owner or occupant) can explicitly express the desired level of system performance in the form of a guarantee. For example, the decision maker could state the system is to be sized

with such a capacity c that there is a 90% chance to satisfy the peak demand D at any given hour. The mathematical representation is $p(c \geq D) = 0.9$ which expresses that the probability that c is equal or larger than D is 0.9. It is straightforward to find the size of the system that meets this guarantee with the CDF. The extra system capacity between the expected peak demand, $E(D)$, and the system capacity, c, is denoted as the system margin $\delta = c - E(D)$.

6.4 A case study

6.4.1 Building description

The Hinman case building is used to demonstrate the proposed framework. Descriptions on the building refer to the Chapter 5. See Figure 6.3 for a photo and an energy model of the building. We consider the building at its design stage for which the primary equipment of a centralized HVAC system is to be sized. In terms of systems

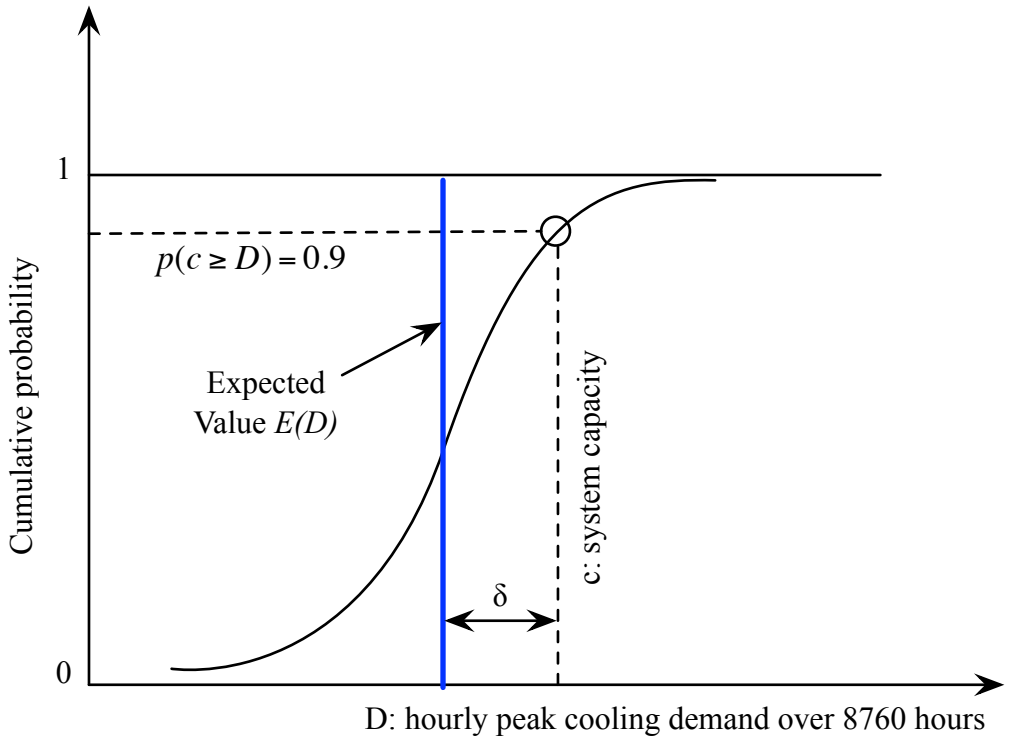


Figure 6.2 Quantify margins based on uncertainty analysis

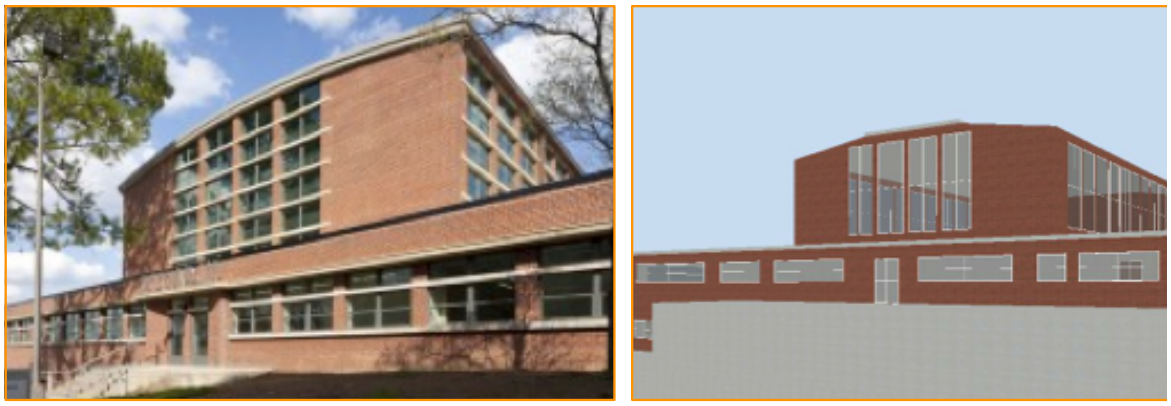


Figure 6.3 Hinman research building; Left: Photo, Right: Design Builder Model

type, we select the VAV (variable air volume) with terminal reheat system. Four air handling units (AHU) are used, i.e. three AHU's for each story and the fourth for the high bay. Air-cooled chillers and gas-fired boilers provide the chilled water to the cooling coils and the hot water to the heating coils in the AHU's.

After the system types are determined, we shall proceed to calculate the system demands to size each component. Here, we use EnergyPlus to carry out the modeling and simulation jobs. We only present the detailed results of the chiller and boiler sizing. It will become clear that the sizing of other equipment, such as an air handling unit, for example, can be conducted in the same manner.

6.4.2 Weather data

We collect actual meteorological years (AMY) from 1982 to 2013, consisting of hourly values regarding temperature, humidity, wind, solar irradiation, etc. Our objective here is to test whether the design day method can give similar results compared with whole-year simulations using multiple AMYs. In the ASHRAE Handbook of Fundamentals, a variety of design days are developed based on the hourly data in the period of 1982 to 2006, which uses the same data source as in our framework. Let us briefly introduce how design days are derived. The annual 0.4% cooling design day and 99.6% heating design day are taken as examples. The hourly dry bulb temperatures from

1982 to 2006 are put together into a single vector, which is then sorted from high to low. The value of dry-bulb temperature at the 0.4% percentile corresponds to the 0.4% cooling design day, and the 99.6% percentile corresponds to the 99.6% heating design day. When scaled to one year, 0.4% of 8760 hours equals 35 hours. Similar to the annual 0.4% cooling design day, annual 1.0, 2.0, and 5.0% design days are also developed. In the handbook, the meanings of these design days are explained as follows: "The design values occur more frequently than the corresponding nominal percentile in some years and less frequently in others. The 0.4, 1.0, 2.0, and 5.0% values are exceeded on average 35, 88, 175, and 438 h per year, respectively, for the period of record."

We take the annual 0.4% cooling design day and 99.6% heating design day as examples to test whether the building demand under the design day conditions can approximate the similar percentile value of the 8760 hourly demands obtained from whole-year simulation using AMYs. To make a fair comparison, we fix all other (model input) parameters at constant nominal values. We then run whole-year hourly simulation from 1982 to 2013. Each year contains about 8760 hours of building cooling load, summarized in a CDF. Figure 6.4 shows the CDF of 8760 hourly cooling demands per year. The limits of y-axis is set from 0.99 to 1 such that the upper quantile of the CDF can be clearly observed.

The combined CDF from 1982 to 2006 is used to compare with the design day that is derived from the same years. The value of building cooling load at the 0.996^{th} quantile of the combined CDF is 322kW. This means that 0.4% of the total hours for the period of 1982 to 2006 has cooling demand that is large or larger than 322 kW. In contrast, when we use the 0.4% cooling design day, the calculated cooling demand is 385 kW. A 63kW difference (i.e., 19.6% higher than 322 kW) is found. Regarding the heating load, Figure 6.5 shows that the value at the 0.996^{th} quantile of the combined CDF is 340kW, whereas it increases to 443kW when using the 99.6% heat design day. The difference is as much as 103kW, i.e., 30.3% higher than 340kW.

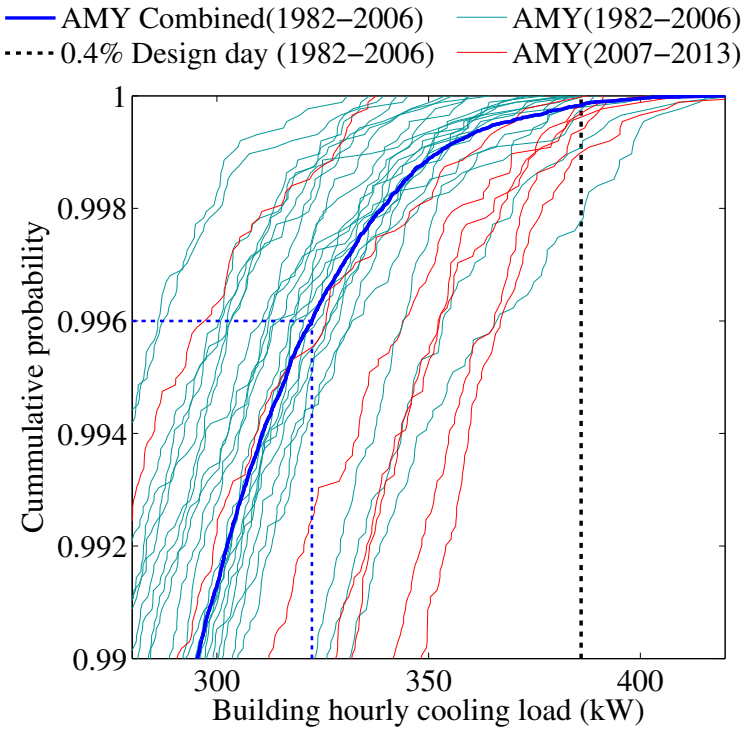


Figure 6.4 Cooling load calculation with multi-year AMYs and design days

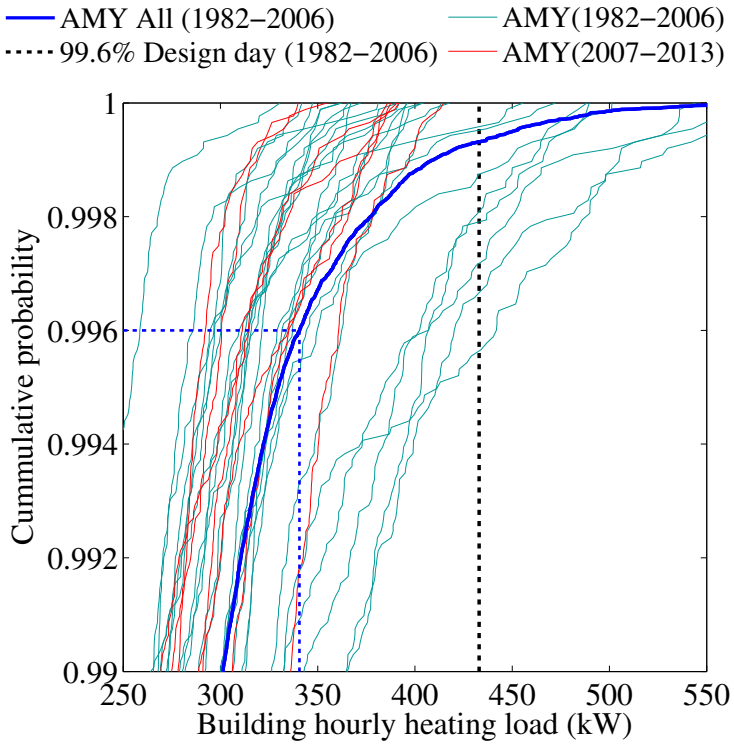


Figure 6.5 Heating load calculation with multi-year AMYs and design days

The simulation results using the more recent AMY, i.e. from 2007 to 2013 is also plotted in Figure 6.4 and Figure 6.5. It shows that both cooling and heating CDF curves of these recent years are well bounded by those from the past years from 1982 to 2006. Therefore, it is suggested that in the period 2007 to 2013 there is no significant warming trend that requires additional consideration of weather uncertainty. It can therefore be concluded that system sizing using recent weather data from 1982 should be enough for short-term predictions. If longer term predictions, e.g to 2050 and beyond, are necessary to size certain system components that have a long service life, there may be a need to account for long-term future trends regarding climate change.

In summary, the 0.4/99.6 cooling and heating demand calculated with the design days are notably higher than the corresponding values using the original full-year weather data. Even though we expected that the cooling and heating design day would lead to overestimation of the cooling and heating load for the reasons explained in section 2, the magnitude of overestimation shown in this case study is somewhat surprising. Moreover, the design-day sizing method was developed under specific practical constraints in terms of: (1) lacking long-term weather measurements and, (2) lacking computational power. Since both constraints have practically been removed by technology advances, we argue that a new framework that uses real, fully correlated weather information at the design stage is warranted to avoid unintentional oversizing.

6.4.3 Uncertainty quantification

We conduct the UQ using a recent developed program, GURA-W, which is integrated with a standard UQ repository for building energy simulation (Lee et al. 2013b). We organize two types of uncertainty, i.e. model form and input parameter uncertainty, and arrange them by spatial scale, going from meteorological weather, to urban microclimate, to building envelop and material, to system components, and to usage scenarios and operations. Table 6.1 summarizes the primary types of UQ.

Table 6.1 Summary of uncertainty quantification

Phenomena/Parameter	Uncertainty quantification	Group
1. Weather	Atlanta AMY from 1982 - 2013	1
2. Microclimate		
Urban heat island effect (°C)	Model form	2
Local wind speed (m/s)	Model form	2
Wind pressure coefficient	Model form	2
Diffuse solar on tilt surfaces (W/m^2)	Model form	2
Ground solar reflectance	Uniform (0.1, 0.3)	2
3. Building envelop		
External convective heat transfer	Bivariate normal	3
$h_{c,ext} = aV_z + b$	$[a; b] \sim \boldsymbol{N}(\mu, \Sigma)$	3
Internal wall convective heat transfer	Bivariate normal	3
$h_{c,wall} = m_w \lvert \Delta T \rvert^{n_w}$	$[m_w; n_w] \sim \boldsymbol{N}(\mu_w, \Sigma_w)$	3
Floor convective heat transfer	Bivariate normal	3
$h_{c,floor} = m_f \lvert \Delta T \rvert^{n_f}$	$[m_f; n_f] \sim \boldsymbol{N}(\mu_f, \Sigma_f)$	3
Ceiling convective heat transfer	Bivariate normal	3
$h_{c,ceil} = m_c \lvert \Delta T \rvert^{n_c}$	$[m_c; n_c] \sim \boldsymbol{N}(\mu_c, \Sigma_c)$	3
Effective leakage area at 4Pa (cm^2/m^2)	Log normal distribution	4
4. Material		
Conductivity (W/m.K)	Relative Normal (1, 10%)	5
Density (kg/m^3)	Relative Normal (1, 10%)	5
Specific heat capacity (kJ/kg.°C)	Relative Normal (1, 10%)	5
Solar absorptance	Relative Normal (1, 10%)	5
Glazing front side solar reflectance	Relative Normal (1, 10%)	5
Glazing back side solar reflectance	Relative Normal (1, 10%)	5
Glazing solar transmittance	Relative Normal (1, 10%)	5
5. Operation		
Lighting peak use (W/m^2)	Relative Uniform (70%, 130%)	6
Plug load peak use (W/m^2)	Relative Uniform (70%, 130%)	6
Occupant density ($\#/m^2$)	Relative Uniform (70%, 130%)	6

Additionally, the uncertainty is classified into six groups as shown in the table. This is used in the group sensitivity analysis that focuses on the combined effect of uncertainty from all parameters within a logical group. We group the uncertainty primarily based on the scale of the uncertainty origins. We separate infiltration from the

envelope scale because infiltration itself can be an influential factor, which has its own physical interpretation.

6.4.4 Probabilistic prediction of cooling and heating demand

With the quantified uncertainty, we now turn to the propagation of uncertainty through simulations in EnergyPlus. Because the weather is a discrete variable with 32 levels (i.e., 1982 – 2013) in our case, to ensure the balance of the weather effect, we use 960 samples to ensure each AMY is used 30 times. In total, we have 960 simulation runs prepared for the following analysis.

We first present the heating and cooling analysis. Since each whole-year simulation contains about 8760 hourly outputs, we use the CDF to visualize the result for each year. Figure 6.6 shows the hourly cooling and heating load for 960 runs. The start point of y-axis is set at 0.99 to focus on the quantile section of our interest. The mean and 95% confidence interval are depicted in the figure as well. We can see that the CDF curves spread over a wide range. This means that given the explicitly quantified

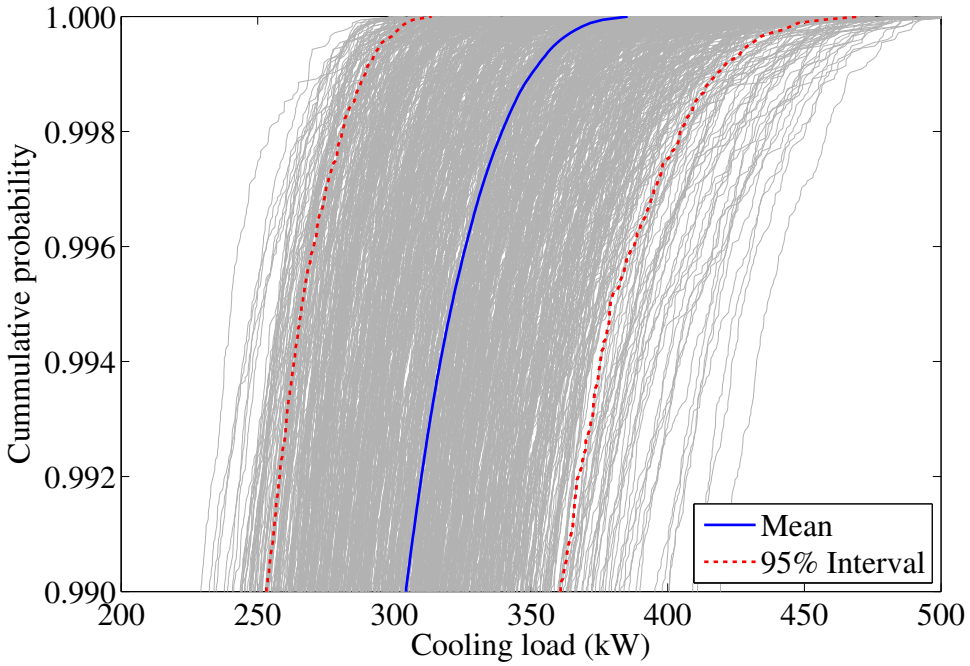


Figure 6.6 Probabilistic prediction of cooling load

uncertainty, the cooling and heating load cannot be predicted with great precision. For practical purposes we could assume that a point prediction on which traditionally methods rely, may at best result in a value somewhere near the mean.

Now we shall discuss how this probabilistic prediction can be used to size the system. Consider we allow 35 unmet hours (i.e., 0.4% of 8760 hours) over a year for either cooling or heating. Figure 6.7 plots the distribution of the cooling load at the 99.6th percentile over the 960 sample runs with an average of 326kW. Note that for any given year in the future, the actual total unmet hours could be more than 35 hours in some situations, and also could be less than 35 hours in other situations because of uncertainty. The expected number of unmet hours is on average 35 taken over all 960 situations we have considered.

What if the building owner is not merely interested in the expected system performance, but wants to control the likelihood that the system fails to provide the required service to be less than a threshold, e.g., 10%. It is indeed a very natural logic that

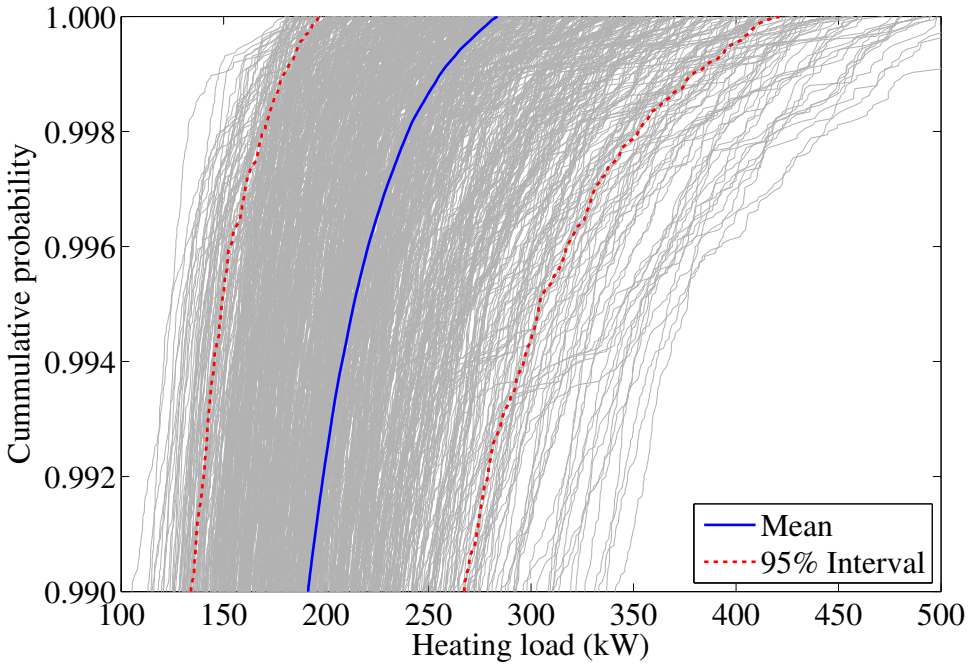


Figure 6.7 Probabilistic prediction of heating load

different building owners differ in their risk attitude towards the system service. Obviously, the traditional sizing framework seems powerless to support such a natural request to manage the risk. In contrast, the proposed framework provides the answer. As shown in the Figure 6.8, the value corresponding to the 90% quantile of the CDF is 374kW. Thus, the cooling component needs to be sized with a marginal capacity of 48kW (compared to the original of 326kW) to meet the objective of the building owner. Obviously the same procedure applies to the sizing of the heating system as shown in Figure 6.9.

6.4.5 Sensitivity analysis

Besides using sensible margins of acceptable unmet hours, it is important to look at system sizing as part of the whole design process where the loads and the systems to satisfy them are considered together. It can for instance be desirable to reduce the uncertainty in system demands to meet the same risk attitude of a building owner. With reduced uncertainty in the loads, one could potentially reduce the size of the system and

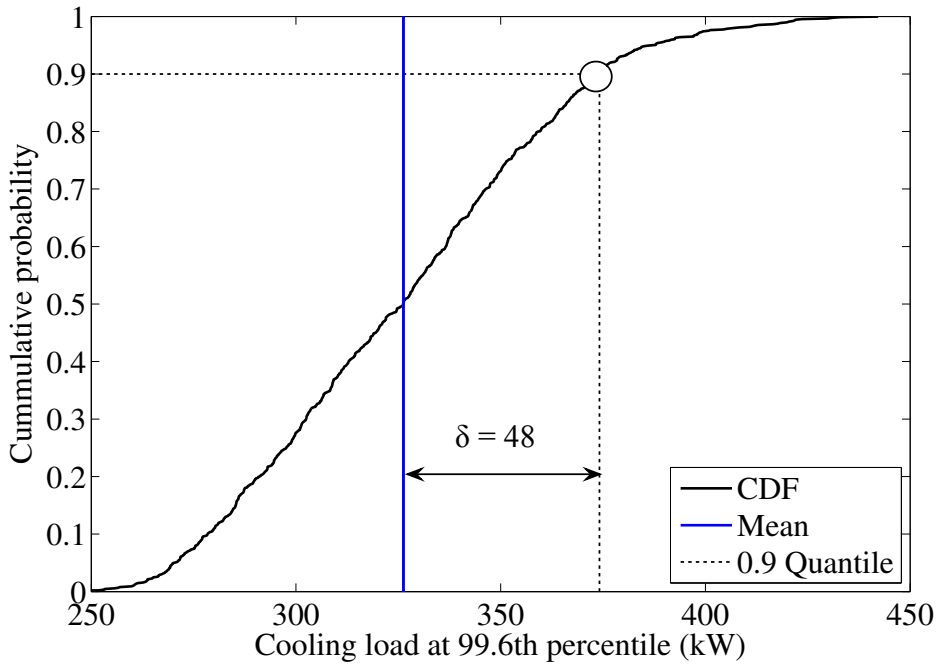


Figure 6.8 Chiller sizing with quantified margins

still meet the same design criterion. Therefore, we will focus our attention on the important sources of uncertainty using sensitivity analysis. Once these sources and their relative importance are known, we will be able to inspect their role in the overall sizing approach.

In total, there are 32 weather years denoted by variable W, and 144 continuous uncertain inputs with index of X1, X2, ⋯, X144. We use the two-step method to calculate the sensitivity index (SI) from the same simulation runs used for UA. The SI for individual parameter with respect to cooling and heating load at the 99.6th percentile is obtained and plotted in Figure 6.10 and Figure 6.11. The R^2 values for cooling and heating are 0.98 and 0.96, respectively. This means that the linear ANOVA model can recover 98 and 96% of the uncertainty in cooling and heating outputs, respectively. The remaining 2 and 3% is due to the interaction and higher-order effect. Moreover, the first two parameters, X131: Occupant density and X134: ELA, account for more than 60% of the output uncertainty in either heating or cooling. This confirms that the building simulation model also conforms to the effect sparsity principle that guiding the

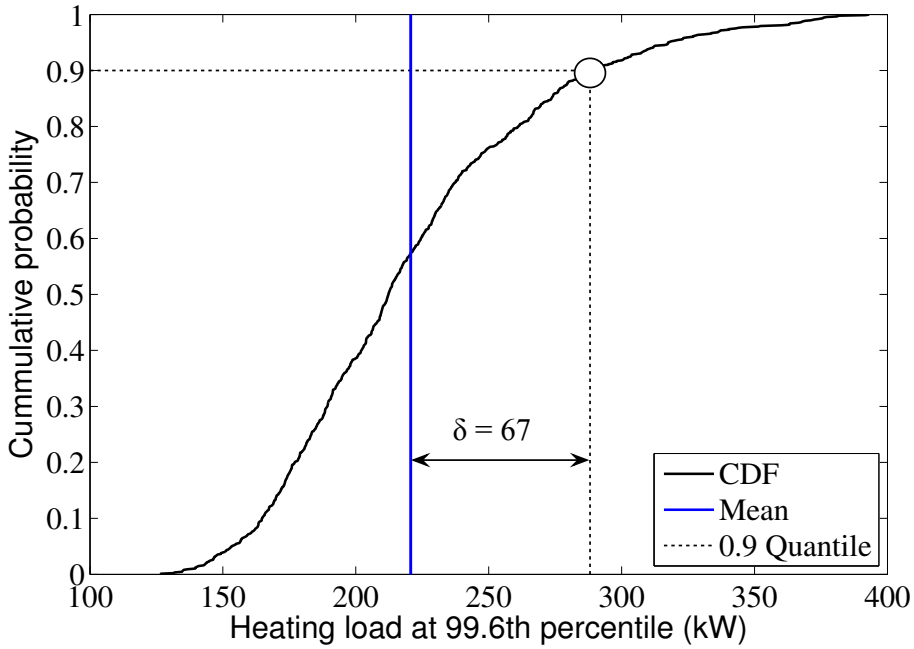


Figure 6.9 Boiler sizing with quantified margins

development of our SA method in Chapter 3. Therefore, the proposed two-step SA method is very efficient and easy to use as a statistical postprocessor of UA to generate informative feedback to the design team.

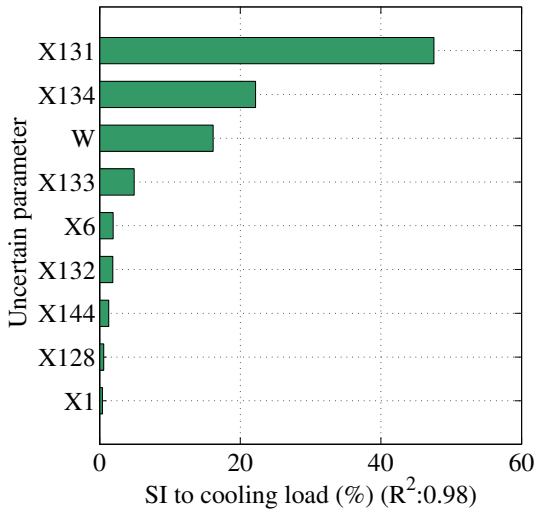


Figure 6.10 Sensitivity indices of uncertainty parameters to cooling load

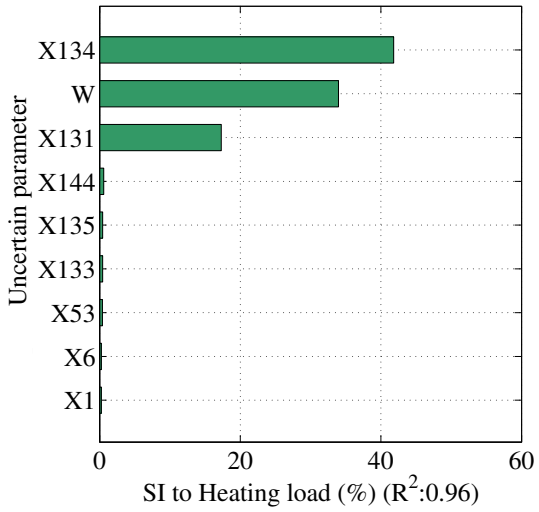


Figure 6.11 Sensitivity indices of uncertainty parameters to heating load

After we obtain SI for each individual parameter, we turn to group SA. The parameters are classified into six groups (Table 6.1). Figure 6.12 and Figure 6.13 show the SI's for the groups, which represent the combined effect of parameters in each group.

For example, Group 6, operational parameters, is the most sensitive group for cooling load uncertainty, whereas Group 4, infiltration, is the most sensitive group for heating load uncertainty, followed by the weather uncertainty. More importantly, the group SA leads to more intuitive interpretation and easier guidance of uncertainty management than individual parameter SI. The group SI can tell the allocation of resources to reduce the uncertainty and estimate the payback in terms of using smaller equipment. In our case,

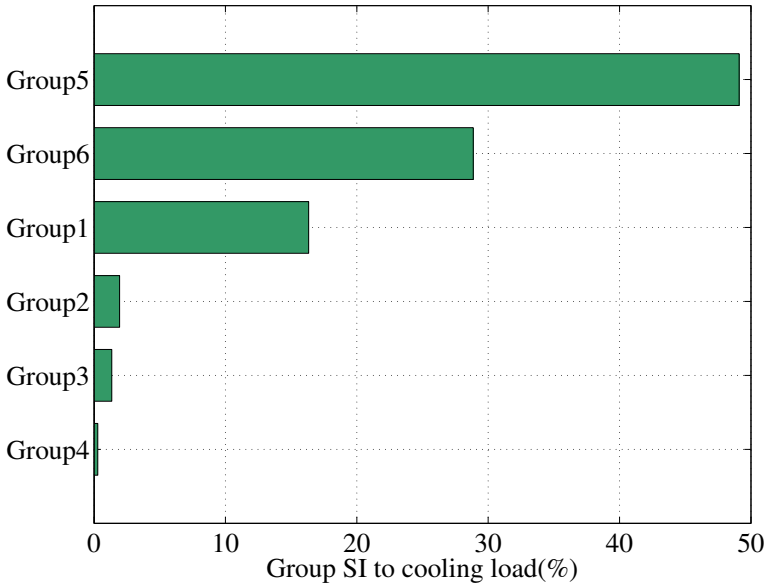


Figure 6.12 Cooling sensitivity indices of groups

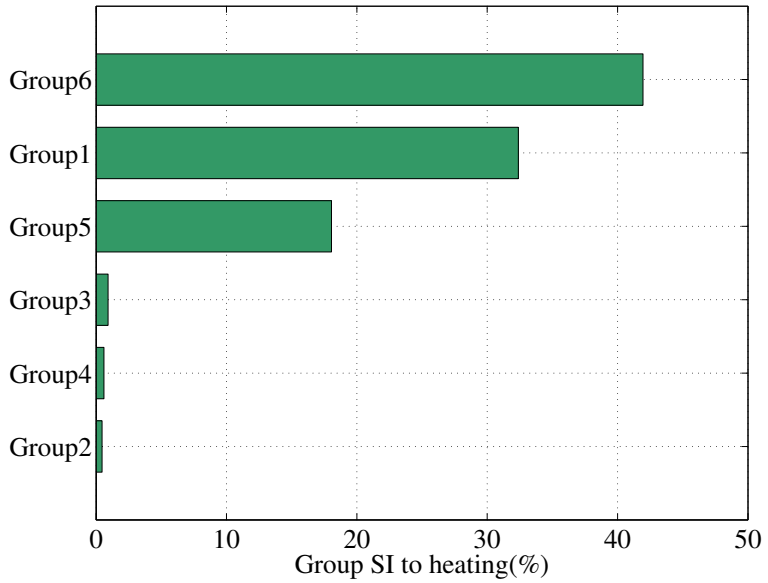


Figure 6.13 Heating sensitivity indices of groups

consider that we have no better knowledge or control over weather uncertainty (Group 1) and operational uncertainty (Group 6). The infiltration uncertainty, Group 4, becomes the clear target for both cooling and heating. A natural follow-up question is whether infiltration control technologies should be deployed, and if so, will the cost be paid off from reducing the size of HVAC system while keeping the same level of system performance? This would typically entail a second-round UA and SA, as part of a continuous and iterative UA process as UQ in certain parameters and parameter groups is adapted by better knowledge, deeper inspection or realization guarantees, such as packaged in performance subcontracts.

6.5 Summary

Existing HVAC sizing methods based on standardized design day methods offer no transparency of the guarantees (in terms of probability) that a system design meets the required level of performance. The level of performance (i.e. meeting peak load in 99.6% of all hours) is only implicitly guaranteed by the definition of the method and choice of safety factor. This leads to undesirable practices of over-engineering.

A new design method based on dynamic simulation with inclusion of quantified uncertainties is proposed. The method is defined through clearly defined steps, using mainstream dynamic simulation and emerging standardized uncertainty repositories for building input and model parameters.

The new design model has three significant advantages, (1) use of real weather data thus removing unwarranted extremes from current design day methods, (2) explicit tracking of the probability that the system meets the 99.6 percentile (or any other target), (3) ability to support risk based sizing to meet stakeholder specified guarantees, (4) support of sensitivity analysis to drive system size reductions by building improvements while satisfying identical system operation guarantees.

CHAPTER 7 CONCLUSIONS AND FUTURE WORK

7.1 Summary and conclusions

Adequately predicting building energy consumption is vital to the success of high-performance building design and retrofit. Building energy models are applied to compliance testing against building energy codes, and the assessment of the saving potential in retrofitting existing buildings. In spite of large investments in model and tool development from the government, research, and industry communities, deterministic model predictions turn out to be anything but assurance that a certain energy performance will be achieved. Many case buildings across commercial and residential sectors confirmed the urgency to investigate the discrepancy between predicted and measured energy uses. If not done well, this performance gap issue will undermine one's confidence in energy-efficient buildings, and thereby the role of building energy efficiency in the national carbon reduction plan.

With the ultimate goal of bridging the energy performance gap, the thesis deals with the problem through achieving three successive research objectives: (1) to build a theoretical basis for understanding the building energy performance gap, (2) to enhance our capability in predicting future building energy use, i.e. closing the building energy performance gap in a model predictive context, and (3) to offer advice to the building industry on uncertainty management strategies. The particular findings can be summarized as follows:

1. We developed a formal framework for closing the building energy performance gap. The inception of the performance gap concept is under deterministic prediction and post-occupant verification contexts, such that the distance of the point prediction with the real-valued observation is interpreted as the performance gap. We argue that using historical data from buildings in operation allows us to observe this

problem, but this does not necessarily mean that the problem can be prevented for future cases. Therefore, we propose to probe the performance gap in a model prediction context when the buildings are still in the design stage, representing the unknown building energy consumption with probability distribution. With the new framework, two pathways towards the closure of the building energy performance gap are identified: (1) enhance model prediction capability, (2) reduce uncertainty.

2. Predicting building energy consumption in a probabilistic way is the key enabler of closing the energy performance gap. The probabilistic prediction is regarded as a second layer on top of the traditionally deterministic prediction, meaning that the fidelity of the underling physical model is just as important as the modeling of the uncertainties. Hence, we choose EnergyPlus as our tool base since it is regarded as the start-of-the-art building energy modeling and simulation package. We organize the uncertainties stemming from five system scales: meteorology, urban, building, system, and occupant with the ambition to conduct exhaustive uncertainty quantification (UQ). At each system scale, we investigate the involved parameters as well as physical phenomena described by different submodels. According to our experience and accessible database, we choose a list of UQ targets that are expected to be influential for building energy predictions, especially at time-aggregated levels. Parameter uncertainties are quantified primarily as parametric probability distributions, considering correlations when necessary. As a relatively new subject in the building energy modeling domain, model form UQ as introduced in this thesis is a methodology contribution of our research. We presented general methods for model form UQ using either high-fidelity models or physical measurements, and both methods involve statistical modeling to describe model form uncertainty which intends to correct the bias in the baseline model prediction in a probabilistic way.

In addition, a UQ repository is evolving, as the generic data and model basis of uncertainty and risk analysis in building energy related applications. In order to make

processing elegant and efficient, we store the results of UQ in an XML based repository. Model form uncertainties are implemented into the source code of EnergyPlus V7.0. The new compiled version with these extensions is the "intrusive" part of the overall UQ repository.

3. We present methods for evaluating the probabilistic predictions by pooling incompatible data from different buildings. Evaluating predictions against the subsequent measurements offers feedbacks to the model development team. This feedback constitutes the information source for augmenting our confidence in using the model for future cases. We adapt methods from other science and engineering fields for the verification of building performance predictions. Starting from a theoretical framework for prediction verification that is general application to either deterministic or probabilistic predictions, we introduce a set of scoring functions for two forms of building energy predictions. Focusing on probabilistic prediction verification, we introduce the Probabilistic Integral Transform (PIT) method to pool incompatible model verification results from different buildings, including the case where a single-real valued observation is paired with a probabilistic prediction for verification purpose. Pooling PITs from different case buildings allows us to evaluate the underling model or modeling procedures that generate the predictions for each case. The results of the PIT method are graphically presented, showing the consistency of probabilistic predictions with the measurements. Additionally, we propose to use the continuous ranked probability score (CRPS) to quantitatively evaluate model prediction. The CRPS is shown to assess model predictions by simultaneously penalizing errors in predicting the means as well as the uncertainties.

4. We show substantial improvement in model prediction capability using six case buildings on Georgia Tech campus. Results from the six case buildings show that compared with the traditional simulation results using EnergyPlus V7.0, the UQ repository with both parameter and model form UQ can substantially reduce prediction

errors. In particular, using the best point estimations of the probabilistic predictions reduces the mean absolute error (MAE) from 44% to 15% and the root mean squared error (RMSE) from 49% to 18% in total annual cooling energy consumption. As for monthly cooling energy consumption, the MAE decreases from 44% to 21% and the RMSE decreases from 53% to 28%. More importantly, the entire probability distributions provided by the UQ repository is statistically verified on the annual aggregated building energy predictions. We also show that model form uncertainty must be quantified to provide high-fidelity probabilistic predictions. The ignorance of the model form uncertainty leads to biased probabilistic predictions. Using CRPS as a measure indicates that our model form UQ at a variety of subsystems reduces the probabilistic prediction error at the whole building-level by 50% and 25% for annual and monthly cooling energy use, respectively.

7.2 Recommendations for future study

The role of building energy modeling and simulation has increased significantly in recent years and will continue to grow. Model prediction capability needs to be continuously enhanced and assured to become more adequate to support many types of decisions. We recommend future studies on the following issues:

1. The UQ repository needs to be enriched and categorized based on the needs of decisions. It is our experience that UQ is most adequate if it is done after a model prediction context is specified with a set of quantities of interest. For different applications, the UQ in some parameters or model forms might be done differently. Also, the current state of the UQ repository focuses more on the demand side than the supply side of the building energy systems. This is the natural tendency in building energy modeling research as there are so many physical processes at the demand side that attract researchers to make better models or quantify their inadequacies as uncertainties.

Exploratory study shows other sources of uncertainty might have equal or even more important impact on outcomes. For example, the uncertainty in our current models of HVAC air supply systems, water supply systems, or control systems is found to be considerable in predicting building performance, especially at hourly or even shorter time steps. Finally, the uncertainty distributions in certain parameters, such as the lighting and plug loads, need to be further refined as more data are available, categorized for different type of buildings, occupants, organizations, etc.

2. The probabilistic prediction is verified for the cooling energy consumption with only six buildings on Georgia Tech campus. More case studies are necessary to confirm our findings across larger sets of buildings of different types and for various energy measures.

3. The thesis focuses on the building energy performance gap resulted from dynamic simulations. Although the developed UQ methods and prediction verification methods still hold for other types of predictions, we have not explored the necessity of dynamic simulation in predicting aggregated energy consumption. Although detailed dynamic simulation offers flexibility in representing building design specifications in computer models, it remains unknown whether the use of sophisticated dynamic simulation should still be the method of choice for routine applications. This becomes an important consideration when one realizes that complex and versatile modeling tools introduce additional uncertainty encapsulated in modeler's choice. The question has pragmatic significance yet has to be researched with enough data and collected under systematic design of experiments.

4. The confounding effect of model form and parameter uncertainty necessitates a new look at data-driven calibration that is still an active research topic and a key technology of (virtual) auditing. It should be realized that data-driven calibration might be subject to seriously biased inference on the values of the parameters in the presence of ignored model form uncertainty that is shown to be quite significant by our cases. This

thesis contributes to the ongoing projects in the building research community, specifically validating data-driven calibration for identifying opportunities of building retrofits. The methods and models developed for prediction verification in the thesis should be equally applicable for evaluating the calibration results of either deterministic or probabilistic form. In addition, the expended UQ repository including the intrusive part will augment the basis of data and tools necessary for improving the efficiency and accuracy of data-driven calibration.

5. The past several years has seen a widespread belief that analytics offers value for building efficiency. New technologies are collecting more data from buildings than ever before. The expansion of data collection, management, and sharing across building types, climates, and countries has become an important theme that receives funds from governments and other organizations. It is thus reasonable to anticipate that data paucity will not be the biggest obstacle in the near future. Meanwhile, it may be critical to research about how to obtain sharper, more timely insights from the analytics, and how the analytics can help us understand what is happening now, what is likely to happen next and what actions should be taken to get the optimal results.

Based on our experience from the thesis related work, we suggest thinking about these questions priori to collecting, cleaning, and reserving the data. It is not hard to find examples in the literature that making data the overriding priorities often fails to delivery the real insights that justify the resources spent on the data. Although that starting from the questions may sound an old-fashioned research approach, we find it is still effective in dealing with large-scale data sets from buildings. Thus, we suggest first defining the questions, which could be either to verify the efficiency of current technologies or to explore new ones, needed to achieve the project goals and then identifying the gaps of information that the data and analytics can bridge.

Another overarching issue of analytics is about model selection. Physical engineering modeling and statistical modeling are two main means to quantitatively

explore the behaviors of systems. The former is built on physical principles, representing the underling mechanism explicitly, whereas the latter excels at dealing with error and noise inherent in the data while using very simple (e.g. linear) approximations to the underling mechanism. The thesis explores the interaction between the two modeling methods and shows promising complementarity between them demonstrated by cases. We believe there are many other questions in building science and engineering domain that can be similarly researched.

REFERENCES

Abushakra, B, A Sreshthaputra, JS Haberl, and DE Claridge. 2001. "Compilation of Diversity Factors and Schedules for Energy and Cooling Load Calculations, Final Report of the ASHRAE Research Project 1093-RP. Energy Systems Laboratory, Texas A&M University.

Ahrens, C. Donald. 2007. *Meteorology today : an introduction to weather, climate, and the environment*. Cengage Learning.

AIAA. 1998. "Guide for the Verification and Validation of Computational Fluid Dynamics Simulations." American Institute of Aeronautics and Astronautics. AIAA-G-077-1998, Reston, VA.

Allaire, Douglas Lawrence. 2009. "Uncertainty Assessment of Complex Models with Application to Aviation Environmental Systems." Massachusetts Institute of Technology.

ASHRAE. 1975. *Subroutine algorithms for heating and cooling loads to determine building energy requirements*, *Energy calculations*. In. New York: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

ASHRAE. 2004. "Standard Method of Test for the Evaluation of Building Energy Analysis Computer Programs." In. Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

ASHRAE. 2007a. "ANSI/ASHRAE/IESNA Standard 90.1-2007: Energy Standard for Buildings Except Low-Rise Residential Buildings " In. Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers

ASHRAE. 2007b. "Standard Method of Test for the Evaluation of Building Energy Analysis Computer Programs." In. Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

ASHRAE. 2009. *ASHRAE handbook fundamentals*. Atlanta, GA: American Society of Heating, Refrigeration and Air-Conditioning Engineers.

ASME. 2009. "Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer." In. The American Society of Mechanical Engineers.

Augenbroe, Godfried. 2012. "EFRI-SEED Annual Project Report: Risk Conscious Design and Retrofit of Buildings for Low Energy." In.: College of Architecture, Georgia Institute of Technology, Atlanta, GA.

Augenbroe, Godfried, Yuna Zhang, Javad Khazaii, Heng Su, Yuming Sun, Benjamin D Lee, and C.F. Jeff Wu. 2013. Implication of the uncoupling of building and HVAC simulation in the presence of parameter uncertainties. Paper presented at the Conference of International Building Performance Simulation Association, Chambery, France.

Barley, C Dennis, M Deru, S Pless, P Torcellini, and S Hayter. 2005. "Procedure for measuring and reporting commercial building energy performance" In. National Renewable Energy Laboratory, Golden, Colorado.

Bedford, T., and Roger M. Cooke. 2001. *Probabilistic risk analysis : foundations and methods*. Cambridge, UK: Cambridge University Press.

Bordass, B, Robert Cohen, and John Field. 2004. Energy performance of non-domestic buildings: closing the credibility gap. Paper presented at the Building Performance Congress. Frankfurt, Germany.

Bordass, Bill, Adrian Leaman, and Paul Ruyssevelt. 2001. "Assessing building performance in use 5: conclusions and implications." *Building Research & Information* 29 (2):144-57.

Box, George EP, and KB Wilson. 1951. "On the experimental attainment of optimum conditions." *Journal of the Royal Statistical Society. Series B (Methodological)* 13 (1):1-45.

BSRIA, UBT. 2009. "The Soft Landings Framework for better briefing, design, handover and building performance in-use." In. Bracknell, UK. BSRIA Ltd.

Burnham, Kenneth P, and David R Anderson. 2002. *Model selection and multi-model inference: a practical information-theoretic approach*: Springer.

Casati, B., L. J. Wilson, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, and S. Mason. 2008. "Forecast verification: current status and future directions." *Meteorological Applications* 15 (1):3-18.

Chen, W., Y. Xiong, K. L. Tsui, and S. Wang. 2008. "A design-driven validation approach using Bayesian prediction models." *Journal of Mechanical Design* 130 (2).

Claridge, D. E., B. Abushakra, J. S. Haberl, and A. Sreshthaputra. 2004. "Electricity Diversity Profiles for Energy Simulation of Office Buildings. " *ASHRAE Transactions* 110.1: 365-377.

Clarke, J. A. 2001. *Energy simulation in building design*. 2nd ed. Oxford ; Boston: Butterworth-Heinemann.

Clarke, J. A., J. Cockroft, S. Conner, J. W. Hand, N. J. Kelly, R. Moore, T. O'Brien, and P. Strachan. 2002. "Simulation-assisted control in building energy management systems." *Energy and Buildings* 34 (9):933-40.

Cole, R. J., and N. S. Sturrock. 1977. "The convective heat exchange at the external surface of buildings." *Building and Environment* 12 (4):207-14.

Coleman, Hugh W., W. Glenn Steele, Hugh W. Coleman, and Wiley InterScience (Online service). 2009. *Experimentation, validation, and uncertainty analysis for engineers*. Hoboken, N.J.: John Wiley & Sons.

Cooke, RogerM. 2013. "Uncertainty analysis comes to integrated assessment models for climate change…and conversely." *Climatic Change* 117 (3):467-79.

Cóstola, D., B. Blocken, M. Ohba, and J. L. M. Hensen. 2010. "Uncertainty in airflow rate calculations due to the use of surface-averaged pressure coefficients." *Energy and Buildings* 42 (6):881-8.

De Wilde, Pieter. 2014. "The gap between predicted and measured energy performance of buildings: A framework for investigation." *Automation in Construction* 41: 40-9.

De Wit, Sten. 2001. "Uncertainty in predictions of thermal comfort in buildings." Delft University of Technology.

De Wit, Sten, and Godfried Augenbroe. 2002. "Analysis of uncertainty in building design evaluations and its implications." *Energy and Buildings* 34 (9):951-8.

DEC. 2014. "Display Energy Certificates in the UK." Accessed May 2.

Demanuele, Christine, Tamsin Tweddell, and Michael Davies. 2010. Bridging the gap between predicted and actual energy performance in schools. Paper presented at the World renewable energy congress XI.

Deru, M, K Field, D Studer, K Benne, B Griffith, P Torcellini, B Liu, M Halverson, D Winiarski, and M Rosenberg. 2011. "US Department of Energy Commercial Reference Building Models of the National Building Stock." In.: National Renewable Energy Laboratory, Golden, CO.

Desogus, Giuseppe, Salvatore Mura, and Roberto Ricciu. 2011. "Comparing different approaches to in situ measurement of building components thermal resistance." *Energy and Buildings* 43 (10):2613-20.

Djunaedy, Ery, Kevin Van den Wymelenberg, Brad Acker, and Harshana Thimmana. 2011. "Oversizing of HVAC system: signatures and penalties." *Energy and Buildings* 43 (2):468-75.

Domínguez-Muñoz, Fernando, Brian Anderson, José M. Cejudo-López, and Antonio Carrillo-Andrés. 2010. "Uncertainty in the thermal conductivity of insulation materials." *Energy and Buildings* 42 (11):2159-68.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. "Least angle regression." *The Annals of statistics* 32 (2):407-99.

EnergyPlus. 2012. "Energyplus Engineering Reference." Accessed 11/26. http://apps1.eere.energy.gov/buildings/energyplus/energyplus_documentation.cfm.

Faraway, Julian J. 2004. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*: CRC press.

Felts, Don, and Patrick Bailey. 2000. The state of affairs—packaged cooling equipment in California. Paper presented at the Proceedings of the 2000 ACEEE Summer Study on Energy Efficiency in Buildings.

Ferson, Scott, Cliff A. Joslyn, Jon C. Helton, William L. Oberkampf, and Kari Sentz. 2004. "Summary from the epistemic uncertainty workshop: consensus amid diversity." *Reliability Engineering & System Safety* 85 (1–3):355-69.

Gneiting, Tilmann. 2011. "Making and Evaluating Point Forecasts." *Journal of the American Statistical Association* 106 (494):746-62.

Gneiting, Tilmann, and Matthias Katzfuss. 2014. "Probabilistic Forecasting." *Annual Review of Statistics and Its Application* 1 (1):125-51.

Gneiting, Tilmann, and Adrian E Raftery. 2007. "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association* 102 (477):359-78.

Griffith, B., N. Long, P. Torcellini, R. Judkoff, D. Crawley, and J. Ryan. 2007. "Assessment of the Technical Potential for Achieving Net Zero-Energy Buildings in the Commercial Sector." In, National Renewable Energy Laboratory. Golden, CO.

Gueymard, Christian. 1987. "An anisotropic solar irradiance model for tilted surfaces and its comparison with selected engineering algorithms." *Solar Energy* 38 (5):367-86.

Guyon, Gilles. 1997. Role of the Model User in Results Obtained From Simulation Software Program. Paper presented at the Building Simulation Conference, Prague, Czech Republic.

Hagishima, Aya, and Jun Tanimoto. 2003. "Field measurements for estimating the convective heat transfer coefficient at building surfaces." *Building and Environment* 38 (7):873-81.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. 2009. *The elements of statistical learning*: Springer.

Hay, John E, and DC McKay. 1988. *Calculation of solar irradiances for inclined surfaces: Verification of models which use hourly and daily data*: International Energy Agency.

Helton, J. C. 2009. "Conceptual and Computational Basis for the Quantification of Margins and Uncertainty." In.: Albuquerque, NML Sandia National Laboratories.

Helton, J. C., and F. J. Davis. 2003. "Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems." *Reliability Engineering and System Safety* 81:23-69.

Helton, J. C., J. D. Johnson, and W. L. Oberkampf. 2004. "An exploration of alternative approaches to the representation of uncertainty in model predictions." *Reliability Engineering & System Safety* 85 (1–3):39-71.

Helton, J. C., and C. J. Sallaberry. 2009. "Computational implementation of sampling-based approaches to the calculation of expected dose in performance assessments for the proposed high-level radioactive waste repository at Yucca Mountain, Nevada." *Reliability Engineering & System Safety* 94 (3):699-721.

Helton, Jon C., Clifford W. Hansen, and Cédric J. Sallaberry. 2012. "Uncertainty and sensitivity analysis in performance assessment for the proposed high-level radioactive waste repository at Yucca Mountain, Nevada." *Reliability Engineering & System Safety* 107:44-63.

Heo, Yeonsook, Godfried Augenbroe, and Ruchi Choudhary. 2012b. "Quantitative risk management for energy retrofit projects. " *Journal of Building Performance and Simulation* 6(4): 257-268.

Heo, Yeonsook, Ruchi Choudhary, and Godfried Augenbroe. 2012. "Calibration of building energy models for retrofit analysis under uncertainty." *Energy and Buildings* 47:550-60.

Hills, R. G., K. J. Dowding, and L. Swiler. 2008. "Thermal challenge problem: Summary." *Computer Methods in Applied Mechanics and Engineering* 197 (29–32):2490-5.

Hills, Richard G., Martin Pilch, Kevin J. Dowding, John Red-Horse, Thomas L. Paez, Ivo Babuška, and Raul Tempone. 2008. "Validation Challenge Workshop." *Computer Methods in Applied Mechanics and Engineering* 197 (29–32):2375-80.

Hu, Huafen. 2009. "Risk-Conscious Design of Off-grid Solar Energy Houses." Georgia Institute of Technology.

Hu, Huafen, and Godfried Augenbroe. 2012. "A stochastic model based energy management system for off-grid solar houses." *Building and Environment* 50:90-103.

Zero Carbon Hub. 2011. "The Performance Challenge: A programme to close the gap between deisgned and as-built performance." Zero Carbon Hub, London, UK.

Hunn, B. D., and D. O. Calafell. 1977. "Determination of Average Ground Reflectivity for Solar Collectors." *Solar Energy* 19 (1):87-9.

IEEE. 1990. "IEEE standard glossary of software engineering terminology." In *IEEE Std*. New York.

Janssen, Hans. 2013. "Monte-Carlo based uncertainty analysis: Sampling efficiency and sampling convergence." *Reliability Engineering & System Safety* 109:123-32.

Jennings, Burgess Hill. 1970. *Environmental engineering; analysis and practice*. Scranton, Pa.,: International Textbook Co.

Jolliffe, I. T., and David B. Stephenson. 2003. *Forecast verification : a practitioner's guide in atmospheric science*. 2nd ed. Chichester, West Sussex ; Hoboken, N.J.: Wiley-Blackwell.

Judkoff, R., D. Wortman, B. O'Doherty, and J. Burch. 2008. "A Methodology for Validating Building Energy Analysis Simulations." In.: National Renewable Energy Laboratory, Golden, CO.

Judkoff, R., D.N. Wortman, and J. Burch. 1983. "Empirical validation using data from the SERI class-A Validation House." In.: Solar Energy Research Institute, Golden, Colorado USA.

Kennedy, Marc C., and Anthony O'Hagan. 2001. "Bayesian Calibration of Computer Models." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63 (3):425-64.

Kimura, Ken'ichi. 1977. *Scientific basis of air conditioning*, *Architectural science series*. London: Applied Science Publishers.

Konikow, Leonard F., and John D. Bredehoeft. 1992. "Ground-water models cannot be validated." *Advances in Water Resources* 15 (1):75-83.

Kumar, Subodh, V. B. Sharma, T. C. Kandpal, and S. C. Mullick. 1997. "Wind induced heat losses from outer cover of solar collectors." *Renewable Energy* 10 (4):613-6.

Laurent, Marie-Hélène, Benoît Allibe, T Oreszczyn, I Hamilton, R Galvin, and C Tigchelaar. 2013. "Back to reality: How domestic energy efficiency policies in four European countries can be improved by using empirical data instead of normative calculation." In.: ECEEE Summer Study Proceedings, Club Belambra Les Criques, France.

Lee, Benjamin D, Yuming Sun, Godfried Augenbroe, and Christiaan J. J Paredis. 2013. Towards Better Prediction of Building Performance: A Workbench to Analyze uncertainty in building simulation. Paper presented at the Conference of International Building Performance Simulation Association, Chambery, France.

Lee, Benjamin D, Yuming Sun, Huafen Hu, Godfried Augenbroe, and Christiaan J. J Paredis. 2012. "A Framework for Generating Stochastic Meteorological Years for Risk-Conscious Design of Buildings." *Proceedings of IBPSA-USA 2012,* Madison Wisconsin.

Li, H., J. Harvey, and A. Kendall. 2012. "Field measurement of albedo for different land cover materials and effects on thermal performance." *Building and Environment* 59: 536-546.

Liu, B, and R Jordan. 1961. "Daily insolation on surfaces tilted towards equator." *Trans. ASHRAE* 53:526-41.

Loutzenhiser, Peter G., Heinrich Manz, Stephan Carl, Hans Simmler, and Gregory M. Maxwell. 2008. "Empirical validations of solar gain models for a glazing unit with exterior and interior blind assemblies." *Energy and Buildings* 40 (3):330-40.

Macdonald, I. A. 2002. "Quantifying the effects of uncertainty in building simulation." University of Strathclyde.

Macdonald, Iain, and Paul Strachan. 2001. "Practical application of uncertainty analysis." *Energy and Buildings* 33 (3):219-27.

Mckay, M. D., R. J. Beckman, and W. J. Conover. 1979. "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code." *Technometrics* 21 (2):239-45.

Menezes, Anna Carolina, Andrew Cripps, Dino Bouchlaghem, and Richard Buswell. 2012. "Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap." *Applied Energy* 97:355-64.

Moon, H. J., and G. Augenbroe. 2007. "Application of probabilistic simulation and Bayesian decision theory in the selection of mold remediation actions." *Building Simulation 2007, Vols 1-3, Proceedings*:912-8.

Morgan, M Granger. 2009. *Best practice approaches for characterizing, communicating and incorporating scientific uncertainty in climate decision making*. DIANE Publishing.

Morgan, M Granger, and David W Keith. 2008. "Improving the way we think about projecting future energy use and emissions of carbon dioxide." *Climatic Change* 90 (3):189-215.

Muneer, Tariq. 2004. *Solar Radiation and Daylight Models*: Routledge.

NBI. 2008a. *Energy performance of LEED for new construction buildings*: New Buildings Institute Washington, DC.

NBI. 2008b. "Energy performance of LEED for new construction buildings." In.: New Buildings Institute, Vancouver.

Nicol, K. 1977. "Energy-Balance of an Exterior Window Surface, Inuvik, Nwt, Canada." *Building and Environment* 12 (4):215-9.

NOAA. 2014. "National Climatic Data Center." Accessed March 10. http://www.ncdc.noaa.gov/cdo-web/.

Noorian, Ali Mohammad, Isaac Moradi, and Gholam Ali Kamali. 2008. "Evaluation of 12 models to estimate hourly diffuse irradiation on inclined surfaces." *Renewable Energy* 33 (6):1406-12.

NRC. 2012. *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*: The National Academies Press.

O'Connor, Patrick, and Andre Kleyner. 2011. *Practical reliability engineering*: John Wiley & Sons.

O'Hagan, Anthony, and Jeremy E. Oakley. 2004. "Probability is perfect, but we can't elicit it perfectly." *Reliability Engineering & System Safety* 85 (1–3):239-48.

Oberkampf, W.L., T.G. Trucano, and C. Hirsch. 2004. "Verification, validation, and predictive capability in computational engineering and physics." *Applied Mechanics Reviews*, *57*(5), 345-384.

Oberkampf, William L., and Christopher J. Roy. 2010. *Verification and validation in scientific computing*. New York: Cambridge University Press.

Palyvos, J. A. 2008. "A survey of wind convection coefficient correlations for building envelope energy systems' modeling." *Applied Thermal Engineering* 28 (8-9):801-8.

Perez, R., P. Ineichen, R. Seals, J. Michalsky, and R. Stewart. 1990. "Modeling daylight availability and irradiance components from direct and global irradiance." *Solar Energy* 44 (5):271-89.

Qian, Peter ZG, and C.F. Jeff Wu. 2008. "Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments." *Technometrics* 50 (2):192-204.

Reda, Ibrahim. 2011. "Method to Calculate Uncertainties in Measuring Shortwave Solar Irradiance Using Thermopile and Semiconductor Solar Radiometers." In, NREL, Golden, CO.

Roache, P.J. 2009. *Fundamentals of verification and validation*: Hermosa publishers.

Roberts, David, Noel Merket, Ben Polly, Mike Heaney, Sean Casey, and Joseph Robertson. 2012. "Assessment of the US Department of Energy's Home Energy Scoring Tool." In, National Renewable Energy Laboratory, Golden, CO.

Rosen, Robert. 1991. *Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life*: Columbia University Press.

Roth, KW, D. Westphalen, MY Feng, P. Llana, and L. Quartararo. 2005. "Energy impact of commercial building controls and performance diagnostics: market characterization, energy impact of building faults and energy savings potential." *Prepared by TAIX LLC for the US Department of Energy. November*.

Roy, Christopher J., and William L. Oberkampf. 2011. "A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing." *Computer Methods in Applied Mechanics and Engineering* 200 (25-28):2131-44.

Ryan, Emily M., and Thomas F. Sanquist. 2012. "Validation of building energy modeling tools under idealized and realistic conditions." *Energy and Buildings* 47:375-82.

Saltelli, Andrea, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. 2008. *Global sensitivity analysis: the primer*: John Wiley & Sons.

Saltelli, Andrea, Stefano Tarantola, and Francesca Campolongo. 2000. "Sensitivity analysis as an ingredient of modeling." *Statistical Science*:377-95.

Santner, Thomas J, Brian J Williams, and William I Notz. 2003. *The design and analysis of computer experiments*: Springer.

Schlesinger, S., R.E. Crosbie, R.E. Gagne, G.S. Innis, CS Lalwani, J. Loch, R.J. Sylvester, R.D. Wright, N. Kheir, and D. Bartos. 1979. "Terminology for model credibility." *Simulation* 32 (3):103-4.

Scofield, John H. 2009. "Do LEED-certified buildings save energy? Not really…." *Energy and Buildings* 41 (12):1386-90.

Seber, George AF, and Alan J Lee. 2012. *Linear regression analysis*. Vol. 936: John Wiley & Sons.

Sharples, S. 1984. "Full-scale measurements of convective energy losses from exterior building surfaces." *Building and Environment* 19 (1):31-9.

Sloughter, J. M., T. Gneiting, and A. E. Raftery. 2010. "Probabilistic Wind Speed Forecasting Using Ensembles and Bayesian Model Averaging." *Journal of the American Statistical Association* 105 (489):25-35.

SRML. "http://solardat.uoregon.edu/SolarData.html."

Stein, Jeff Ross, and Alan Meier. 2000. "Accuracy of home energy rating systems." *Energy* 25 (4):339-54.

Strachan, P. 1993. "Model validation using the PASSYS Test cells." *Building and Environment* 28 (2):153-65.

Sun, Yuming, and Godfried Augenbroe. 2014. "Urban heat island effect on energy application studies of office buildings." *Energy and Buildings* 77:171-9.

Sun, Yuming, Yeonsook Heo, Matthias H. Y. Tan, Huizhi Xie, C. F. Jeff Wu, and Godfried Augenbroe. 2014. "Uncertainty quantification of microclimate variables in building energy models." *Journal of Building Performance Simulation* 7 (1):17-32.

Sun, Yuming, Heng Su, C. F. Jeff Wu, and Godfried Augenbroe. 2014. "Quantification of model form uncertainty in the calculation of solar diffuse irradiation on inclined surfaces for building energy simulation." *Journal of Building Performance Simulation*. To Appear. DOI:10.1080/19401493.2014.914247

Sunikka-Blank, Minna, and Ray Galvin. 2012. "Introducing the prebound effect: the gap between performance and actual energy consumption." *Building Research & Information* 40 (3):260-73.

Taki, A. H., and D. L. Loveday. 1996. "External convection coefficients for framed rectangular elements on building facades." *Energy and Buildings* 24 (2):147-54.

Thornton, BA, MI Rosenberg, EE Richman, W Wang, Y Xie, J Zhang, H Cho, VV Mendon, RA Athalye, and B Liu. 2011. "Achieving the 30% Goal: Energy and Cost Savings Analysis of ASHRAE Standard 90.1-2010." In, edited by Pacific Northwest National Laboratory. Richland.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*:267-88.

TM54. 2013. "Evaluating operational energy performance of buildings at the design stage." CIBSE, Lavenham Press.

Torcellini, Paul Allen, Shanti Pless, Michael Deru, Brent Griffith, Nick Long, and Ron Judkoff. 2006. "Lessons learned from case studies of six high-performance buildings." In.: National Renewable Energy Laboratory, Golden, CO.

Turner, C., M. Frankel, and U.S.G.B. Council. 2008. *Energy performance of LEED for new construction buildings*: New Buildings Institute, Washington DC.

U.S.EPA. 2008. "National Action Plan for Energy Efficiency Vision for 2025: A Framework for Change." U.S. Environmental Protection Agency, Washington, D.C.

Walker, Warren E, Poul Harremoës, Jan Rotmans, Jeroen P van der Sluijs, Marjolein BA van Asselt, P Janssen, and MP Krayer von Krauss. 2003. "Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support." *Integrated assessment* 4 (1):5-17.

Wang, Liping, Paul Mathew, and Xiufeng Pang. 2012. "Uncertainties in energy consumption introduced by building operations and weather for a medium-size office building." *Energy and Buildings* 53:152-8.

Wang, Richard Y, and Diane M Strong. 1996. "Beyond accuracy: What data quality means to data consumers." *Journal of Management Information Systems* 12 (4):5-33.

Williamson, T. J. 2010. "Predicting building performance: the ethics of computer simulation." *Building Research & Information* 38 (4):401-10.

Wu, C.F. Jeff, and Michael Hamada. 2009. *Experiments: Planning, Analysis, and Optimization*. Hoboken, New Jersey: John Wiley.

Yang, Hongxing, Lin Lu, and Wei Zhou. 2007. "A novel optimization sizing model for hybrid solar-wind power generation system." *Solar Energy* 81 (1):76-84.

Yazdanian, M., and J. H. Klems. 1994. Measurement of the exterior convective film coefficient for windows in low-rise buildings. Paper presented at the Proceedings of the ASHRAE Winter Meeting, January 23, 1994 - January 26, 1994, New Orleans, LA, USA.